A Research Of Seismic Data Reconstruction Based On Conditional Constraint Diffusion Model

Fei Deng[1], Shuang Wang[1], Xuben Wang[1], and Peng Fang[2]

[1]Chengdu University of Technology
[2]Institute of geology and geophysics, Chinese academy of science

June 7, 2023

Abstract

Reconstruction of complete seismic data is a crucial step in seismic data processing, which has seen the application of various convolutional neural networks (CNNs). These CNNs typically establish a direct mapping function between input and output data. In contrast, diffusion models which learn the feature distribution of the data, have shown promise in enhancing the accuracy and generalization capabilities of predictions by capturing the distribution of output data. However, diffusion models lack constraints based on input data. In order to use the diffusion model for seismic data interpolation, our study introduces conditional constraints to control the interpolation results of diffusion models based on input data. Furthermore, we improving the sampling process of the diffusion model to ensure higher consistency between the interpolation results and the existing data. Experimental results conducted on synthetic and field datasets demonstrate that our method outperforms existing methods in terms of achieving more accurate interpolation results.

A Research Of Seismic Data Reconstruction Based On Conditional Constraint Diffusion Model

Fei Deng[1], Shuang Wang[2], Xuben Wang[3]and Peng Fang[4]

[1]College of Computer Science and Cyber Security, Chengdu University of Technology, Chengdu, China
[2]College of Computer Science and Cyber Security, Chengdu University of Technology, Chengdu, China
[3]College of Geophysics,Chengdu University of Technology, Chengdu, China
[4]Chinese academy of science, Institute of geology and geophysics, Peking, China

Key Points:

- Introducing diffusion models for seismic data reconstruction.
- Conditional constraints are employed to constrain the interpolation results according to the input data.
- Improving the sampling process to ensure greater consistency between the interpolation results and the original data.

Corresponding author: Shuang Wang, wangs@stu.cdut.edu.cn

Abstract

Reconstruction of complete seismic data is a crucial step in seismic data processing, which has seen the application of various convolutional neural networks (CNNs). These CNNs typically establish a direct mapping function between input and output data. In contrast, diffusion models which learn the feature distribution of the data, have shown promise in enhancing the accuracy and generalization capabilities of predictions by capturing the distribution of output data. However, diffusion models lack constraints based on input data. In order to use the diffusion model for seismic data interpolation, our study introduces conditional constraints to control the interpolation results of diffusion models based on input data. Furthermore, we improving the sampling process of the diffusion model to ensure higher consistency between the interpolation results and the existing data. Experimental results conducted on synthetic and field datasets demonstrate that our method outperforms existing methods in terms of achieving more accurate interpolation results.

Plain Language Summary

Due to natural or economic constraints, acquired prestack seismic data often exhibits missing traces, making it essential to reconstruct complete seismic data during the data processing stage. While various convolutional neural networks with distinct structures have been used for seismic missing traces interpolation, their direct mapping relationship between input datas and output datas can lead to deviations between the interpolation results and the ground truth. Alternatively, diffusion models, as a novel deep learning model, exhibit higher generative accuracy and generalization ability by learning data distribution. However, as pure generative models, diffusion models do not utilize existing data to guide the generation of unknown data. In order to use the diffusion model for seismic data interpolation, we introduce conditional constraints to control the interpolation results based on the input data and improve the sampling process to maintain greater consistency between the interpolation results and the existing data. Experimental results conducted on both synthetic and field datasets demonstrate that our proposed method yields more accurate interpolation results compared to discriminative-based methods.

1 Introduction

In seismic exploration, seismic data plays a pivotal role as the foundation for analysis and interpretation. However, there are instances where seismic acquisition systems cannot be deployed in certain areas due to factors such as economic or natural constraints, as well as geographical or physical limitations (Kuijpers et al., 2021). Consequently, this leads to the occurrence of consecutive missing traces in the prestack seismic data (Wei et al., 2021; Pawelec et al., 2021). The presence of missing traces severely impacts the subsequent processing and analysis of seismic data, underscoring the need for a crucial step: the reconstruction of complete seismic data.

The methods for interpolating and reconstructing irregular seismic data can be divided into two main categories:traditional interpolation based on the mathematical or physical properties of the data (Zhou & Han, 2018), and deep learning-based methods that utilize neural networks to interpolate irregular data (Jia & Ma, 2017; Park et al., 2021). Methods based on mathematical or physical properties, such as the frequency-space (FX) prediction filtering method (Naghizadeh & Sacchi, 2009) and the projection onto convex sets (POCS) algorithm based on curvelet transform (Yang et al., 2012), are not dataset-specific. However, they are not as effective in handling complex field data and continuous large gaps. Therefore, they are often used as alternative approaches. On the other hand, deep learning-based methods

are not limited by data complexity and can effectively capture the features between traces (Pan et al., 2020), resulting in better reconstruction outcomes. For example, ResNet-based data interpolation method proposed by B. Wang et al. (2019), U-Net network used by Chai et al. (2020) for seismic data reconstruction, convolutional autoencoders proposed by Y. Wang et al. (2020) for interpolating missing traces, the reconstruction network combining deep learning with traditional methods introduced by Zhang et al. (2020), multistage U-Net trained by He et al. (2021) achieving certain results in interpolating low amplitude missing components, and the attention mechanisms incorporated by Yu and Wu (2021) with a hybrid loss function to further improve the reconstruction capability of the U-Net network.

Convolutional discriminative neural networks are capable of directly obtaining predictive outputs through the network, establishing a direct mapping relationship between input and output datas. However, interpolating practical data poses certain challenges, particularly when dealing with limited samples or continuous large gaps in the data traces. To address this issue, we propose a seismic data interpolation method based on a diffusion model (J. Song et al., 2020; Rombach et al., 2022). This method leverages the ability to learn the distribution (Dhariwal & Nichol, 2021) of existing seismic data, enabling it to achieve superior results compared to existing methods. It demonstrates effectiveness in interpolating both high and low amplitude missing components, as well as large gap continuous missing traces and small gap random missing traces.

This paper presents a novel deep learning paradigm, the diffusion model, for seismic data reconstruction. It outlines the architecture and mathematical principles of the diffusion model, which originally produces unconstrained results that are not correlated with the distribution of existing data, making it unsuitable for seismic data reconstruction. To address this issue, we propose the following improvements and contributions:

1.To guide the generation of data based on the input seismic data, we incorporate conditional constraints into the diffusion model.

2.To avoid generating conflicting data distributions with the original data distribution, we improve the sampling process by constraining the generation process through reverse diffusion iterations that sample from the given data.

The comparative experimental results on synthetic and field datasets demonstrate the superiority of our method over existing approaches in terms of achieving more accurate interpolation results. Furthermore, the diffusion model exhibits superior generative accuracy and enhanced generalization ability by learning the underlying data distribution(Dhariwal & Nichol, 2021). Consequently, our network enables the generalization to increasingly complex missing scenarios during the inference process.

2 Diffusion model

Diffusion model is a probabilistic generative model that learns the encoding distribution through the encoding process and then uses a neural network to reverse the encoding process to obtain the decoding distribution. The distribution itself is not the target data, so the reparameterization trick (D. P. Kingma & Welling, 2013) is employed to sample deterministic target data from the decoding distribution. This approach effectively avoids encoding distortion and reduces the deviation between the generated data and the ground truth.

2.1 Training Process:

During the training process, the diffusion model defines a forward encoding process. This process gradually encodes a real-space vector x_0 into a latent-space vector x_T over T encoding steps, with x_T following a Gaussian distribution

$x_T \sim \mathcal{N}(0, \mathbf{I})$. In the diffusion model, based on the Langevin dynamics, a custom variance schedule can be used to stabilize the encoding process (Y. Song & Ermon, 2019). The encoding process from step $t-1$ to step t can be defined as follows:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(\mu_t, \beta_t \mathbf{I}) \tag{1}$$

The value of β_t is obtained from a predefined variance table and typically linearly increases from 0.0001 to 0.002. μ_t represents the mean, and according to Nichol and Dhariwal (2021), $\mu_t = \sqrt{1-\beta_t} x_{t-1}$. Therefore, equation (1) can be rewritten as:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t} x_{t-1}, \beta_t \mathbf{I}) \tag{2}$$

According to Ho et al. (2020), the encoding formula (3) from step 0 to step t can be derived from equation (2):

$$q(x_t \mid x_0) = \mathcal{N}(\sqrt{\overline{\alpha}_t} x_0, (1-\overline{\alpha}_t)\mathbf{I}) \tag{3}$$

Simultaneously, reparameterization can be used to obtain x_t.

$$x_t = \sqrt{\overline{\alpha}_t} x_0 + \epsilon_t \sqrt{(1-\overline{\alpha}_t)}, \epsilon_t \sim \mathcal{N}(0, \mathbf{I}) \tag{4}$$

where $\alpha_t = 1 - \beta_t, \overline{\alpha}_t = \prod_{s=0}^{t} \alpha_s$. ϵ_t is sampled from a Gaussian distribution.

The diffusion model is trained to reverse this process, modelling predicted by a neural network, aiming to obtain the data distribution of the step t-1, denoted as $p_\theta(x_{t-1} \mid x_t)$ as shown in Equation (5). In the diffusion model, p_θ is also a Gaussian distribution (Sohl-Dickstein et al., 2015), so the network needs to estimate the mean $\mu_\theta(x_t, t)$ and variance $\beta_\theta(x_t, t)$ of the distribution.

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(\mu_\theta(x_t, t), \beta_\theta(x_t, t)) \tag{5}$$

To facilitate model training, $\mu_\theta(x_t, t)$ can be further expressed as:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}} \epsilon_\theta(x_t, t)) \tag{6}$$

Expressing $\beta_\theta(x_t, t)$ as:

$$\beta_\theta(x_t, t) = \exp(\epsilon_\theta(x_t, t) \log \beta_t + (1 - \epsilon_\theta(x_t, t)) \log \tilde{\beta}_t) \tag{7}$$

Where $\tilde{\beta}_t = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t} \beta_t$. Both $\mu_\theta(x_t, t)$ and $\beta_\theta(x_t, t)$ are functions of $\epsilon_\theta(x_t, t)$. Therefore, the network only needs to estimate $\epsilon_\theta(x_t, t)$.

To train the network model, considering the variational lower bound (D. Kingma et al., 2021), we can derive the loss function L_{vlb} for the network:

$$L_{vlb} = \mathbb{E}_q[\underbrace{D_{KL}(q(x_T \mid x_0) \parallel p(x_T))}_{L_T} + \sum_{t>1} \underbrace{D_{KL}(q(x_t \mid x_{t-1}, x_0) \parallel p_\theta(x_{t-1} \mid x_t))}_{L_{t-1}} - \underbrace{\log p_\theta(x_0 \mid x_1)}_{L_0}] \tag{8}$$

The diffusion model randomly selects the step t for training during network training process. Therefore, in one training process, only the L_{t-1} loss in the above equation needs to be considered. According to Ho et al. (2020), a simplified loss function can be further derived as shown in Equation (9), where ϵ_t is given by Equation (4).

$$L_{simple} = E_{t, x_0, \epsilon_t}[\|\epsilon_t - \epsilon_\theta(x_t, t)\|^2] \tag{9}$$

2.2 Generation Process:

To generate real-space vectors, the diffusion model iteratively decodes a randomly sampled vector x_T from the T-dimensional latent space, ultimately obtaining a vector x_0 in the real space. The decoding process at step t is as follows:

Using a trained neural network model to predict the mean $\mu_\theta(x_t, t)$ and variance $\beta_\theta(x_t, t)$ of the data distribution at the step t-1, obtaining the data distribution $p_\theta(x_{t-1} \mid x_t)$.

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(\mu_\theta(x_t, t), \beta_\theta(x_t, t)) \tag{10}$$

By utilizing the reparameterization, obtain x_{t-1}.

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t, t)) + \epsilon\sqrt{\beta_\theta(x_t, t)}, \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{11}$$

3 Method

The diffusion model described in the previous section cannot be directly used for seismic data reconstruction. The diffusion model is a purely generative model that can only generate vectors in the real space by sampling from the latent space, once it is trained. However, in seismic data reconstruction, the original data provided by the user must be used for reconstruction, rather than generating randomly. In this section, an improved diffusion model will be proposed to address this issue.

3.1 Resampling

The goal of seismic data reconstruction is to generate unknown traces based on known traces. However, the original diffusion model does not establish a direct link between the generated traces and the known traces, thereby failing to ensure that the distribution of the generated traces aligns with that of the known traces. We use the property that diffusion model naturally aims to generate consistent structural to solve this problem (Lugmayr et al., 2022).

During sampling, the entire seismic data is represented as x, the unknown part is represented as $m \odot x$, and the known part is represented as $(1 - m) \odot x$. From equation (10), it can be observed that each sample x_{t-1} only depends on x_t. Therefore, it is possible to modify the known part $(1-m)\odot x_{t-1}$ of x_{t-1} while maintaining the corresponding distribution. According to (3) and (10), we can obtain:

$$x_{t-1}^{known} \sim \mathcal{N}(\sqrt{\overline{\alpha}_{t-1}}x_0, (1 - \overline{\alpha}_{t-1})\mathbf{I}) \tag{12a}$$

$$x_{t-1}^{unknown} \sim \mathcal{N}(\mu_\theta(x_t, t), \beta_\theta(x_t, t)\mathbf{I}) \tag{12b}$$

$$x_{t-1} = (1 - m) \odot x_{t-1}^{known} + m \odot x_{t-1}^{unknown} \tag{12c}$$

Encode x_{t-1} into x_t using equation (1), at which x_t contains information from the known data, establishing a certain connection between the known and unknown data, reducing data conflicts. Then, obtain x_{t-1} from this x_t using equation (11), and repeat this process.

3.2 Correction

Resampling is used to establish a connection between the known data and the generated data. However, there is a possibility that the reconstructed result may exhibit a distribution similar to the ground truth. During the iterative decoding process of the diffusion model, if the selected vector x_T coincides with the one obtained by encoding the ground truth into the Tth latent space, the decoded vectors in the real space can be considered as the ground truth. However, in practice, the original diffusion model randomly selects the vector x_T, making it unlikely for the decoded vectors to represent the ground truth. Finding the corresponding x_T for the ground truth is particularly challenging, especially in seismic data interpolation where the ground truth itself is uncertain. Therefore, we propose an iterative correction method that gradually approaches the ground truth by incorporating self-supervision constraints in each iterative sampling step. In each step t, the constraint

Algorithm 1 Seismic data reconstruction algorithm.

for $t = T, \ldots, 1$ **do**
  for $u = 1, \ldots, U$ **do**
    $\epsilon \sim \mathcal{N}(0, I)$ if $t > 1$,else $\epsilon = 0$
    $x_{t-1}^{known} = \sqrt{\overline{\alpha}_t} x_0 + \epsilon\sqrt{(1 - \overline{\alpha}_t)})$
    $z \sim \mathcal{N}(0, I)$ if $t > 1$,else $z = 0$
    $x_{t-1}^{unknown} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t, t)) + z\sqrt{\beta_\theta(x_t, t)}$
    $x_{t-1} = (1 - m) \odot x_{t-1}^{known} + m \odot x_{t-1}^{unknown}$
    if $u < U$ and $t > 1$ **then**
      $x_t \sim \mathcal{N}(\sqrt{1 - \beta_{t-1}} x_{t-1}, \beta_{t-1}\mathbf{I})$
    end if
  end for
end for
return x_0

encourages the sampled x_{t-1} to be closer to the representation of the ground truth in the $(t - 1)$th latent space vector, thereby facilitating self-correction within the model. By performing T iterations of correction, the reconstructed vectors in the real space are compelled to approximate the ground truth.

To incorporate self-supervision constraints into the training process of the diffusion model, equation (5) is rewritten as follows:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(\mu_\theta(x_t, y, t), \beta_\theta(x_t, y, t)) \tag{13}$$

Where $\mu_\theta(x_t, y, t)$ is defined as:

$$\mu_\theta(x_t, y, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1 - \overline{\alpha}_t}}\epsilon_\theta(x_t, y, t)) \tag{14}$$

$\beta_\theta(x_t, y, t)$ is represented as:

$$\beta_\theta(x_t, y, t) = \exp(\epsilon_\theta(x_t, y, t) \log \beta_t + (1 - \epsilon_\theta(x_t, y, t)) \log \tilde{\beta}_t) \tag{15}$$

Where $\tilde{\beta}_t = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t$. The network is modified to estimate $\epsilon_\theta(x_t, y, t)$. The variational lower bound loss function L_{vlb} is rewritten as:

$$L_{vlb} = \mathbb{E}_q[\underbrace{D_{KL}(q(x_T \mid x_0) \parallel p(x_T))}_{L_T} + \sum_{t>1} \underbrace{D_{KL}(q(x_t \mid x_{t-1,x_0}) \parallel p_\theta(x_{t-1} \mid x_t, y))}_{L_{t-1}} - \underbrace{\log p_\theta(x_0 \mid x_1, y)}_{L_0}] \tag{16}$$

Based on L_{t-1}, the simplified loss function is rewritten as:

$$L_{simple} = E_{t,x_0,\epsilon_t}[\|\epsilon_t - \epsilon_\theta(x_t, y, t)\|^2] \tag{17}$$

In the generation process, the decoding process at step t is changed as follows:

Using a trained neural network model to predict the mean $\mu_\theta(x_t, y, t)$ and variance $\beta_\theta(x_t, y, t)$ of the data distribution at the step t-1, obtaining the data distribution $p_\theta(x_{t-1} \mid x_t)$.

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(\mu_\theta(x_t, y, t), \beta_\theta(x_t, y, t)) \tag{18}$$

Using the reparameterization, we obtain x_{t-1}:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1 - \overline{\alpha}_t}}\epsilon_\theta(x_t, y, t)) + \epsilon\sqrt{\beta(x_t, y, t)}, \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{19}$$

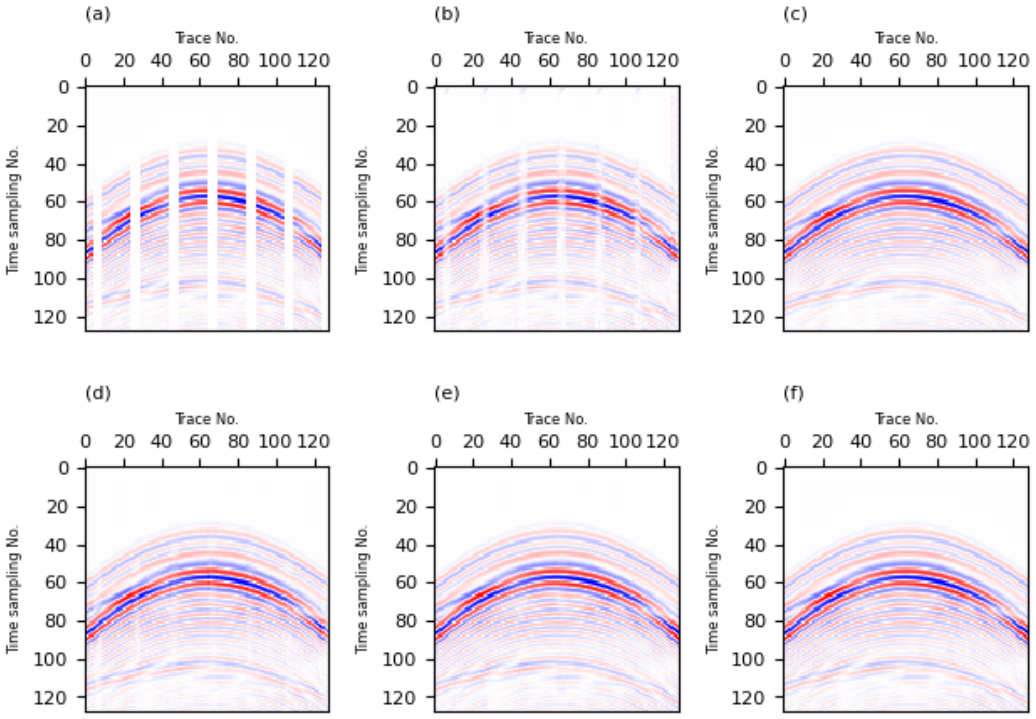

Figure 1. The reconstruction results of different networks for small gap missing traces in the synthetic dataset: (a) Interpolated data, (b) CAE, (c) POCSCNN, (d) ANet, (e) Diffusion model, (f) Ground truth.

4 Experiments

4.1 Synthetic Data

To assess the effectiveness of the proposed method, we conducted experiments on a synthetic dataset using the publicly available Society of Exploration Geophysicists (SEG) C3 dataset. This dataset consists of 45 shots sampled at an 8 ms rate, with each shot containing a receiver grid of size 201 $\times$ 201 and 625 samples per trace.

A total of 1800 patches were selected, out of which 1260 patches were used for training, 360 patches for validation and 180 patches for testing. The value of T for the forward process was set to 1000, and the number of resampling steps was set to 250.

In addition, three different network models were selected for comparative testing, including CAE (Y. Wang et al., 2020), POCSCNN (Zhang et al., 2020), and ANet (Yu & Wu, 2021). Following the methods described in the paper, these models were trained to their optimal states and then compared.

Fig. 1 shows the reconstruction results of the four network models for small gap missing traces. In the data, 28% of the traces were intentionally set to 0 to represent the missing traces, which were distributed in seven locations with each location accounting for 4% of the data. The results demonstrate that CAE did not perform well in the task of reconstruction, POCSCNN exhibited relatively satisfactory results but introduced certain biases, and ANet achieved slightly improved results while still exhibiting some data biases. Conversely, our proposed method yielded the most plausible and reasonable results.

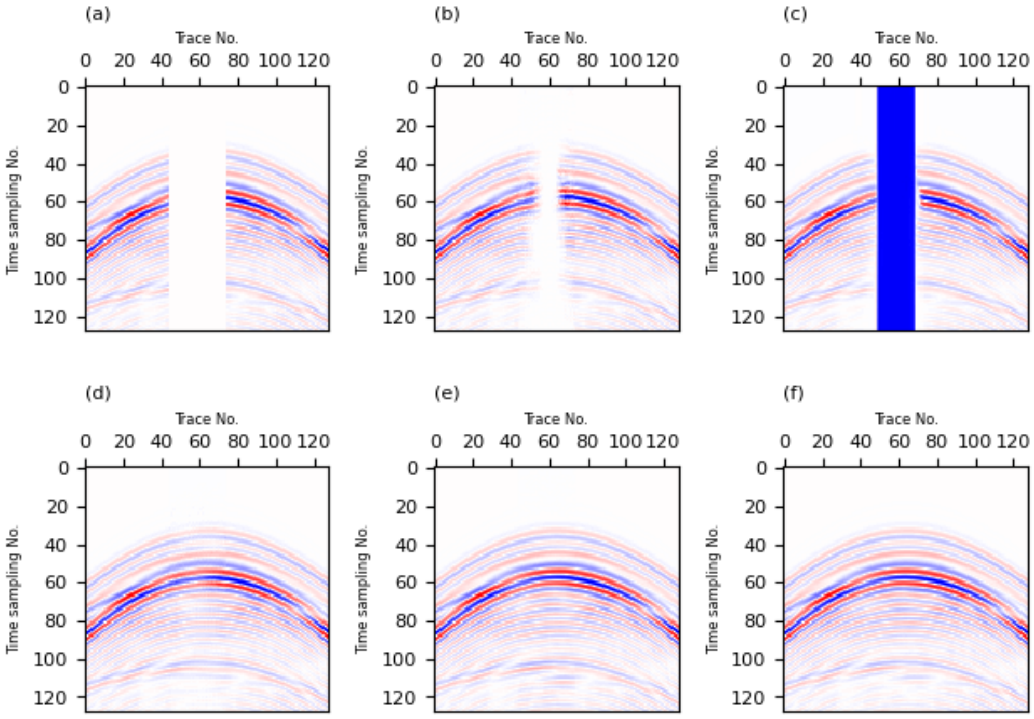

Figure 2. The reconstruction results of different networks for large gap missing traces in the synthetic dataset: (a) Interpolated data, (b) CAE, (c) POCSCNN, (d) ANet, (e) Diffusion model, (f) Ground truth.

To ensure an accurate assessment of the reconstruction results, three commonly employed metrics were employed. Specifically, the Mean Squared Error (MSE), Mean Absolute Error (MAE), and Structural Similarity (SSIM) (Huang et al., 2022) were computed to quantify the disparities between the reconstructed data and the ground truth. The SSIM metric was employed to gauge the resemblance between the two datasets, with values ranging from 0 to 1. A higher SSIM value indicates a greater likeness between the datasets. The comparison of the four network models (see Table S1 in Supporting Information S1) revealing that our proposed method outperforms the other methods in terms of these metrics, demonstrating superior performance.

To evaluate the reconstruction performance of the diffusion model in the context of large gap missing traces, 25% of the consecutive traces in the data were intentionally set to 0 to represent the missing traces. The results were compared with the other three models, as shown in Fig. 2. It can be observed that CAE and POCSCNN performed the worst, with CAE only reconstructing a portion of the traces near the known part, and POCSCNN even experiencing failure in reconstruction. ANet lost some details and had slightly inferior performance compared to the method proposed in this paper. The comparison results of the four networks (see Table S2 in Supporting Information S1) show that our method still exhibited the best performance.

4.2 Field Data

To assess the effectiveness of our method on field data, we conducted experiments on the Mobil Avo Viking Graben Line 12 field dataset and compared it with

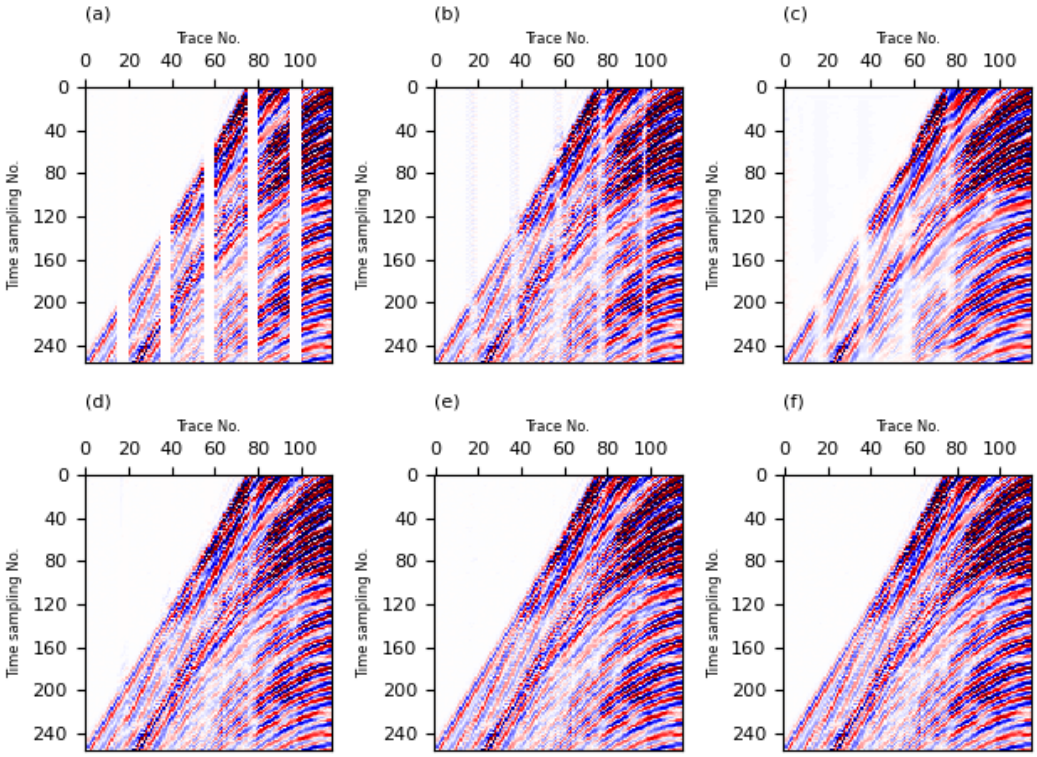

Figure 3. The reconstruction results of different networks for small gap missing traces in the field dataset: (a) Interpolated data, (b) CAE, (c) POCSCNN, (d) ANet, (e) Diffusion model, (f) Ground truth.

three other models. A total of 1000 patches were selected, with 700 patches allocated for training, 200 patches for validation and 100 patches for testing. The value of T for the forward process was set to 1000, and the number of resampling steps was set to 250.

Fig. 3 shows the reconstruction results of the four network models for small gap missing traces in the field dataset. In the data, 20% of the traces were intentionally set to 0 to represent the missing traces, which were distributed in five locations with each location accounting for 4% of the data. It is evident that CAE did not perform well in the reconstruction task, POCSCNN yielded slightly improved results but introduced certain biases, and ANet approached the correct reconstruction but still exhibited some data biases. In contrast, our proposed method produced the most reasonable results.

Calculates the MSE, MAE, and SSIM between the reconstructed data and the ground truth (see Table S3 in Supporting Information S1), it can be observed that our method outperforms the other methods significantly in these metrics, demonstrating superior performance.

To evaluate the reconstruction performance of the diffusion model on large gap missing traces in the field dataset, 25% of the continuous traces in the data were set to 0 as missing traces. A comparison was made with other three models, and the results are shown in Fig. 4. It can be observed that CAE and POCSCNN performed the worst. CAE only reconstructed partial traces near the known part, while POCSCNN even failed to reconstruct. ANet missed some details and had slightly worse performance compared to our method. The comparison of the four networks (see Table S4 in Supporting Information S1), show that our method still exhibited

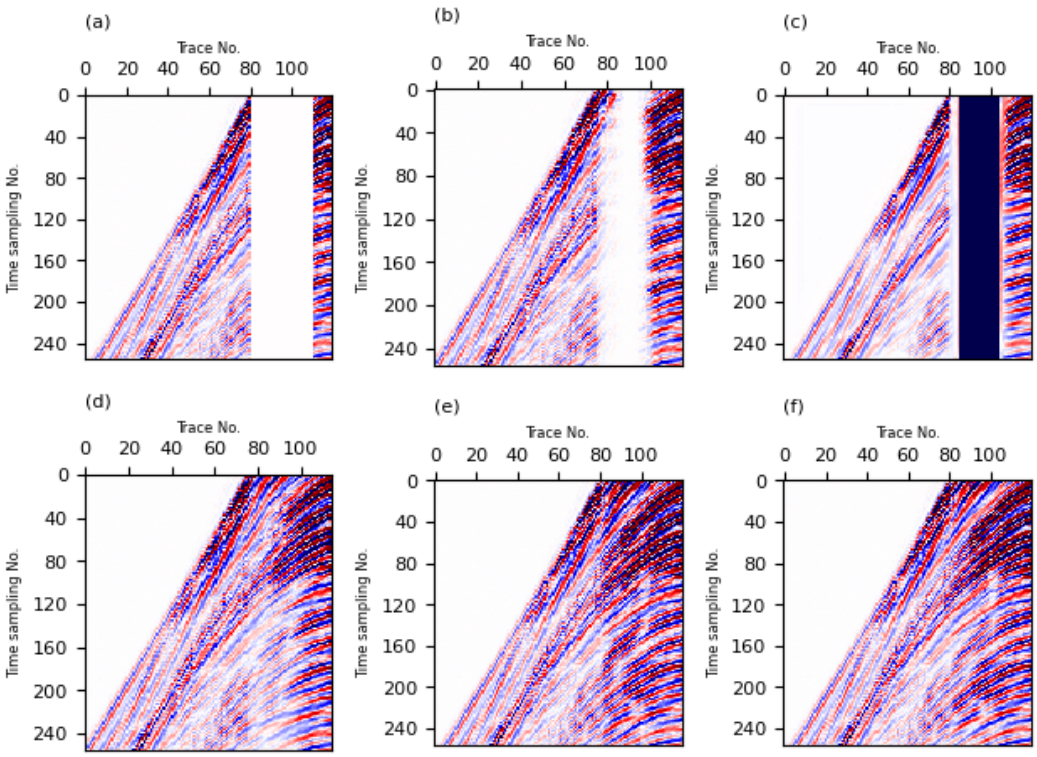

Figure 4. The reconstruction results of different networks for large gap missing traces in the field dataset: (a) Interpolated data, (b) CAE, (c) POCSCNN, (d) ANet, (e) Diffusion model, (f) Ground truth.

superior performance. The above experiments thoroughly validate the effectiveness and applicability of the proposed method in this study.

5 Conclusions

This paper presents a constrained diffusion model for seismic data interpolation and utilize Resampling to impose additional constraints by sampling from given data during the reverse diffusion iterations. This marks the first successful application of the diffusion model in seismic data reconstruction. By learning the distribution of existing seismic data, this method effectively mitigates substantial deviations between generated data and ground truth, which are caused by encoding distortions in traditional convolutional discriminative networks. Comparative experimental results on synthetic and field datasets substantiate that our proposed method achieves more accurate interpolation results compared to existing methods. Additionally, the diffusion model exhibits superior generative accuracy and enhanced generalization ability by learning the data distribution, enabling our network generalizes to more complexity missing scenario during the inference period.

6 Open Research

The observation of Society of Exploration Geophysicists (SEG) C3 dataset (Aminzadeh et al., 1997) of this study are available at `https://wiki.seg.org/wiki/SEG_C3_45_shot`.

The observation of Mobil Avo Viking Graben Line 12 field dataset (Keys & Foster, 1998) of this study are available at `https://wiki.seg.org/wiki/Mobil_AVO_viking_graben_line_12`.

Acknowledgments

The authors sincerely thank Professors Fei Deng for the advice for this study, thanks Peifan Jiang and Bin Wang for the advice.

References

Aminzadeh, F., Brac, J., & Kunz, T. (1997). 3d salt and overthrust models: Presented at the seg. *EAGE Modeling Series*(1).

Chai, X., Gu, H., Li, F., Duan, H., Hu, X., & Lin, K. (2020). Deep learning for irregularly and regularly missing data reconstruction. *Scientific reports*, *10*(1), 3302.

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, *34*, 8780–8794.

He, T., Wu, B., & Zhu, X. (2021). Seismic data consecutively missing trace interpolation based on multistage neural network training process. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, *33*, 6840–6851.

Huang, H., Wang, T., Cheng, J., Xiong, Y., Wang, C., & Geng, J. (2022). Self-supervised deep learning to reconstruct seismic data with consecutively missing traces. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–14.

Jia, Y., & Ma, J. (2017). What can machine learning do for seismic data processing? an interpolation application. *Geophysics*, *82*(3), V163–V177.

Keys, R. G., & Foster, D. J. (1998). A data set for evaluating and comparing seismic inversion methods. *Comparison of seismic inversion methods on a single real data set*, 1–12.

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

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

Kuijpers, D., Vasconcelos, I., & Putzky, P. (2021). Reconstructing missing seismic data using deep learning. *arXiv preprint arXiv:2101.09554*.

Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., & Van Gool, L. (2022). Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 11461–11471).

Naghizadeh, M., & Sacchi, M. (2009). fx adaptive seismic-trace interpolation: Geophysics, 74. *V9–V16*.

Nichol, A. Q., & Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *International conference on machine learning* (pp. 8162–8171).

Pan, S., Chen, K., Chen, J., Qin, Z., Cui, Q., & Li, J. (2020). A partial convolution-based deep-learning network for seismic data regularization1. *Computers & Geosciences*, *145*, 104609.

Park, J., Choi, J., Jee Seol, S., Byun, J., & Kim, Y. (2021). A method for adequate selection of training data sets to reconstruct seismic data using a convolutional u-net. *Geophysics*, *86*(5), V375–V388.

Pawelec, I., Wakin, M., & Sava, P. (2021). Missing trace reconstruction for 2d land seismic data with randomized sparse sampling. *Geophysics*, *86*(3), P25–P36.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the*

ieee/cvf conference on computer vision and pattern recognition (pp. 10684–10695).

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

Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y., & Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, *32*.

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

Wang, Y., Wang, B., Tu, N., & Geng, J. (2020). Seismic trace interpolation for irregularly spatial sampled data using convolutional autoencodercae-based seismic trace interpolation. *Geophysics*, *85*(2), V119–V130.

Wei, Q., Li, X., & Song, M. (2021). Reconstruction of irregular missing seismic data using conditional generative adversarial networks. *Geophysics*, *86*(6), V471–V488.

Yang, P., Gao, J., & Chen, W. (2012). Curvelet-based pocs interpolation of nonuniformly sampled seismic records. *Journal of Applied Geophysics*, *79*, 90–99.

Yu, J., & Wu, B. (2021). Attention and hybrid loss guided deep learning for consecutively missing seismic data reconstruction. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–8.

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

Zhou, Y., & Han, C. (2018). Seismic data restoration based on the grassmannian rank-one update subspace estimation method. *Journal of Applied Geophysics*, *159*, 731–741.

A Research Of Seismic Data Reconstruction Based On Conditional Constraint Diffusion Model

Fei Deng[1], Shuang Wang[2], Xuben Wang[3]and Peng Fang[4]

[1]College of Computer Science and Cyber Security, Chengdu University of Technology, Chengdu, China
[2]College of Computer Science and Cyber Security, Chengdu University of Technology, Chengdu, China
[3]College of Geophysics,Chengdu University of Technology, Chengdu, China
[4]Chinese academy of science, Institute of geology and geophysics, Peking, China

Key Points:

- Introducing diffusion models for seismic data reconstruction.
- Conditional constraints are employed to constrain the interpolation results according to the input data.
- Improving the sampling process to ensure greater consistency between the interpolation results and the original data.

Corresponding author: Shuang Wang, wangs@stu.cdut.edu.cn

Abstract

Reconstruction of complete seismic data is a crucial step in seismic data processing, which has seen the application of various convolutional neural networks (CNNs). These CNNs typically establish a direct mapping function between input and output data. In contrast, diffusion models which learn the feature distribution of the data, have shown promise in enhancing the accuracy and generalization capabilities of predictions by capturing the distribution of output data. However, diffusion models lack constraints based on input data. In order to use the diffusion model for seismic data interpolation, our study introduces conditional constraints to control the interpolation results of diffusion models based on input data. Furthermore, we improving the sampling process of the diffusion model to ensure higher consistency between the interpolation results and the existing data. Experimental results conducted on synthetic and field datasets demonstrate that our method outperforms existing methods in terms of achieving more accurate interpolation results.

Plain Language Summary

Due to natural or economic constraints, acquired prestack seismic data often exhibits missing traces, making it essential to reconstruct complete seismic data during the data processing stage. While various convolutional neural networks with distinct structures have been used for seismic missing traces interpolation, their direct mapping relationship between input datas and output datas can lead to deviations between the interpolation results and the ground truth. Alternatively, diffusion models, as a novel deep learning model, exhibit higher generative accuracy and generalization ability by learning data distribution. However, as pure generative models, diffusion models do not utilize existing data to guide the generation of unknown data. In order to use the diffusion model for seismic data interpolation, we introduce conditional constraints to control the interpolation results based on the input data and improve the sampling process to maintain greater consistency between the interpolation results and the existing data. Experimental results conducted on both synthetic and field datasets demonstrate that our proposed method yields more accurate interpolation results compared to discriminative-based methods.

1 Introduction

In seismic exploration, seismic data plays a pivotal role as the foundation for analysis and interpretation. However, there are instances where seismic acquisition systems cannot be deployed in certain areas due to factors such as economic or natural constraints, as well as geographical or physical limitations (Kuijpers et al., 2021). Consequently, this leads to the occurrence of consecutive missing traces in the prestack seismic data (Wei et al., 2021; Pawelec et al., 2021). The presence of missing traces severely impacts the subsequent processing and analysis of seismic data, underscoring the need for a crucial step: the reconstruction of complete seismic data.

The methods for interpolating and reconstructing irregular seismic data can be divided into two main categories:traditional interpolation based on the mathematical or physical properties of the data (Zhou & Han, 2018), and deep learning-based methods that utilize neural networks to interpolate irregular data (Jia & Ma, 2017; Park et al., 2021). Methods based on mathematical or physical properties, such as the frequency-space (FX) prediction filtering method (Naghizadeh & Sacchi, 2009) and the projection onto convex sets (POCS) algorithm based on curvelet transform (Yang et al., 2012), are not dataset-specific. However, they are not as effective in handling complex field data and continuous large gaps. Therefore, they are often used as alternative approaches. On the other hand, deep learning-based methods

are not limited by data complexity and can effectively capture the features between traces (Pan et al., 2020), resulting in better reconstruction outcomes. For example, ResNet-based data interpolation method proposed by B. Wang et al. (2019), U-Net network used by Chai et al. (2020) for seismic data reconstruction, convolutional autoencoders proposed by Y. Wang et al. (2020) for interpolating missing traces, the reconstruction network combining deep learning with traditional methods introduced by Zhang et al. (2020), multistage U-Net trained by He et al. (2021) achieving certain results in interpolating low amplitude missing components, and the attention mechanisms incorporated by Yu and Wu (2021) with a hybrid loss function to further improve the reconstruction capability of the U-Net network.

Convolutional discriminative neural networks are capable of directly obtaining predictive outputs through the network, establishing a direct mapping relationship between input and output datas. However, interpolating practical data poses certain challenges, particularly when dealing with limited samples or continuous large gaps in the data traces. To address this issue, we propose a seismic data interpolation method based on a diffusion model (J. Song et al., 2020; Rombach et al., 2022). This method leverages the ability to learn the distribution (Dhariwal & Nichol, 2021) of existing seismic data, enabling it to achieve superior results compared to existing methods. It demonstrates effectiveness in interpolating both high and low amplitude missing components, as well as large gap continuous missing traces and small gap random missing traces.

This paper presents a novel deep learning paradigm, the diffusion model, for seismic data reconstruction. It outlines the architecture and mathematical principles of the diffusion model, which originally produces unconstrained results that are not correlated with the distribution of existing data, making it unsuitable for seismic data reconstruction. To address this issue, we propose the following improvements and contributions:

1.To guide the generation of data based on the input seismic data, we incorporate conditional constraints into the diffusion model.

2.To avoid generating conflicting data distributions with the original data distribution, we improve the sampling process by constraining the generation process through reverse diffusion iterations that sample from the given data.

The comparative experimental results on synthetic and field datasets demonstrate the superiority of our method over existing approaches in terms of achieving more accurate interpolation results. Furthermore, the diffusion model exhibits superior generative accuracy and enhanced generalization ability by learning the underlying data distribution(Dhariwal & Nichol, 2021). Consequently, our network enables the generalization to increasingly complex missing scenarios during the inference process.

2 Diffusion model

Diffusion model is a probabilistic generative model that learns the encoding distribution through the encoding process and then uses a neural network to reverse the encoding process to obtain the decoding distribution. The distribution itself is not the target data, so the reparameterization trick (D. P. Kingma & Welling, 2013) is employed to sample deterministic target data from the decoding distribution. This approach effectively avoids encoding distortion and reduces the deviation between the generated data and the ground truth.

2.1 Training Process:

During the training process, the diffusion model defines a forward encoding process. This process gradually encodes a real-space vector x_0 into a latent-space vector x_T over T encoding steps, with x_T following a Gaussian distribution

$x_T \sim \mathcal{N}(0, \mathbf{I})$. In the diffusion model, based on the Langevin dynamics, a custom variance schedule can be used to stabilize the encoding process (Y. Song & Ermon, 2019). The encoding process from step $t-1$ to step t can be defined as follows:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(\mu_t, \beta_t \mathbf{I}) \tag{1}$$

The value of β_t is obtained from a predefined variance table and typically linearly increases from 0.0001 to 0.002. μ_t represents the mean, and according to Nichol and Dhariwal (2021), $\mu_t = \sqrt{1-\beta_t} x_{t-1}$. Therefore, equation (1) can be rewritten as:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t} x_{t-1}, \beta_t \mathbf{I}) \tag{2}$$

According to Ho et al. (2020), the encoding formula (3) from step 0 to step t can be derived from equation (2):

$$q(x_t \mid x_0) = \mathcal{N}(\sqrt{\overline{\alpha}_t} x_0, (1-\overline{\alpha}_t)\mathbf{I}) \tag{3}$$

Simultaneously, reparameterization can be used to obtain x_t.

$$x_t = \sqrt{\overline{\alpha}_t} x_0 + \epsilon_t \sqrt{(1-\overline{\alpha}_t)}, \epsilon_t \sim \mathcal{N}(0, \mathbf{I}) \tag{4}$$

where $\alpha_t = 1 - \beta_t, \overline{\alpha}_t = \prod_{s=0}^{t} \alpha_s$. ϵ_t is sampled from a Gaussian distribution.

The diffusion model is trained to reverse this process, modelling predicted by a neural network, aiming to obtain the data distribution of the step t-1, denoted as $p_\theta(x_{t-1} \mid x_t)$ as shown in Equation (5). In the diffusion model, p_θ is also a Gaussian distribution (Sohl-Dickstein et al., 2015), so the network needs to estimate the mean $\mu_\theta(x_t, t)$ and variance $\beta_\theta(x_t, t)$ of the distribution.

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(\mu_\theta(x_t, t), \beta_\theta(x_t, t)) \tag{5}$$

To facilitate model training, $\mu_\theta(x_t, t)$ can be further expressed as:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}} \epsilon_\theta(x_t, t)) \tag{6}$$

Expressing $\beta_\theta(x_t, t)$ as:

$$\beta_\theta(x_t, t) = \exp(\epsilon_\theta(x_t, t) \log \beta_t + (1 - \epsilon_\theta(x_t, t)) \log \tilde{\beta}_t) \tag{7}$$

Where $\tilde{\beta}_t = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t} \beta_t$. Both $\mu_\theta(x_t, t)$ and $\beta_\theta(x_t, t)$ are functions of $\epsilon_\theta(x_t, t)$. Therefore, the network only needs to estimate $\epsilon_\theta(x_t, t)$.

To train the network model, considering the variational lower bound (D. Kingma et al., 2021), we can derive the loss function L_{vlb} for the network:

$$L_{vlb} = \mathbb{E}_q[\underbrace{D_{KL}(q(x_T \mid x_0) \parallel p(x_T))}_{L_T} + \sum_{t>1} \underbrace{D_{KL}(q(x_t \mid x_{t-1}, x_0) \parallel p_\theta(x_{t-1} \mid x_t))}_{L_{t-1}} - \underbrace{\log p_\theta(x_0 \mid x_1)}_{L_0}] \tag{8}$$

The diffusion model randomly selects the step t for training during network training process. Therefore, in one training process, only the L_{t-1} loss in the above equation needs to be considered. According to Ho et al. (2020), a simplified loss function can be further derived as shown in Equation (9), where ϵ_t is given by Equation (4).

$$L_{simple} = E_{t, x_0, \epsilon_t}[\|\epsilon_t - \epsilon_\theta(x_t, t)\|^2] \tag{9}$$

2.2 Generation Process:

To generate real-space vectors, the diffusion model iteratively decodes a randomly sampled vector x_T from the T-dimensional latent space, ultimately obtaining a vector x_0 in the real space. The decoding process at step t is as follows:

Using a trained neural network model to predict the mean $\mu_\theta(x_t, t)$ and variance $\beta_\theta(x_t, t)$ of the data distribution at the step t-1, obtaining the data distribution $p_\theta(x_{t-1} \mid x_t)$.

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(\mu_\theta(x_t, t), \beta_\theta(x_t, t)) \tag{10}$$

By utilizing the reparameterization, obtain x_{t-1}.

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t, t)) + \epsilon\sqrt{\beta_\theta(x_t, t)}, \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{11}$$

3 Method

The diffusion model described in the previous section cannot be directly used for seismic data reconstruction. The diffusion model is a purely generative model that can only generate vectors in the real space by sampling from the latent space, once it is trained. However, in seismic data reconstruction, the original data provided by the user must be used for reconstruction, rather than generating randomly. In this section, an improved diffusion model will be proposed to address this issue.

3.1 Resampling

The goal of seismic data reconstruction is to generate unknown traces based on known traces. However, the original diffusion model does not establish a direct link between the generated traces and the known traces, thereby failing to ensure that the distribution of the generated traces aligns with that of the known traces. We use the property that diffusion model naturally aims to generate consistent structural to solve this problem (Lugmayr et al., 2022).

During sampling, the entire seismic data is represented as x, the unknown part is represented as $m \odot x$, and the known part is represented as $(1 - m) \odot x$. From equation (10), it can be observed that each sample x_{t-1} only depends on x_t. Therefore, it is possible to modify the known part $(1-m)\odot x_{t-1}$ of x_{t-1} while maintaining the corresponding distribution. According to (3) and (10), we can obtain:

$$x_{t-1}^{known} \sim \mathcal{N}(\sqrt{\overline{\alpha}_{t-1}}x_0, (1 - \overline{\alpha}_{t-1})\mathbf{I}) \tag{12a}$$

$$x_{t-1}^{unknown} \sim \mathcal{N}(\mu_\theta(x_t, t), \beta_\theta(x_t, t)\mathbf{I}) \tag{12b}$$

$$x_{t-1} = (1 - m) \odot x_{t-1}^{known} + m \odot x_{t-1}^{unknown} \tag{12c}$$

Encode x_{t-1} into x_t using equation (1), at which x_t contains information from the known data, establishing a certain connection between the known and unknown data, reducing data conflicts. Then, obtain x_{t-1} from this x_t using equation (11), and repeat this process.

3.2 Correction

Resampling is used to establish a connection between the known data and the generated data. However, there is a possibility that the reconstructed result may exhibit a distribution similar to the ground truth. During the iterative decoding process of the diffusion model, if the selected vector x_T coincides with the one obtained by encoding the ground truth into the Tth latent space, the decoded vectors in the real space can be considered as the ground truth. However, in practice, the original diffusion model randomly selects the vector x_T, making it unlikely for the decoded vectors to represent the ground truth. Finding the corresponding x_T for the ground truth is particularly challenging, especially in seismic data interpolation where the ground truth itself is uncertain. Therefore, we propose an iterative correction method that gradually approaches the ground truth by incorporating self-supervision constraints in each iterative sampling step. In each step t, the constraint

Algorithm 1 Seismic data reconstruction algorithm.

for $t = T, \ldots, 1$ **do**
  for $u = 1, \ldots, U$ **do**
    $\epsilon \sim \mathcal{N}(0, I)$ if $t > 1$,else $\epsilon = 0$
    $x_{t-1}^{known} = \sqrt{\overline{\alpha}_t} x_0 + \epsilon\sqrt{(1 - \overline{\alpha}_t)})$
    $z \sim \mathcal{N}(0, I)$ if $t > 1$,else $z = 0$
    $x_{t-1}^{unknown} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t, t)) + z\sqrt{\beta_\theta(x_t, t)}$
    $x_{t-1} = (1 - m) \odot x_{t-1}^{known} + m \odot x_{t-1}^{unknown}$
    if $u < U$ and $t > 1$ **then**
      $x_t \sim \mathcal{N}(\sqrt{1 - \beta_{t-1}} x_{t-1}, \beta_{t-1}\mathbf{I})$
    end if
  end for
end for
return x_0

encourages the sampled x_{t-1} to be closer to the representation of the ground truth in the $(t - 1)$th latent space vector, thereby facilitating self-correction within the model. By performing T iterations of correction, the reconstructed vectors in the real space are compelled to approximate the ground truth.

To incorporate self-supervision constraints into the training process of the diffusion model, equation (5) is rewritten as follows:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(\mu_\theta(x_t, y, t), \beta_\theta(x_t, y, t)) \tag{13}$$

Where $\mu_\theta(x_t, y, t)$ is defined as:

$$\mu_\theta(x_t, y, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1 - \overline{\alpha}_t}}\epsilon_\theta(x_t, y, t)) \tag{14}$$

$\beta_\theta(x_t, y, t)$ is represented as:

$$\beta_\theta(x_t, y, t) = \exp(\epsilon_\theta(x_t, y, t) \log \beta_t + (1 - \epsilon_\theta(x_t, y, t)) \log \tilde{\beta}_t) \tag{15}$$

Where $\tilde{\beta}_t = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t$. The network is modified to estimate $\epsilon_\theta(x_t, y, t)$. The variational lower bound loss function L_{vlb} is rewritten as:

$$L_{vlb} = \mathbb{E}_q[\underbrace{D_{KL}(q(x_T \mid x_0) \parallel p(x_T))}_{L_T} + \sum_{t>1} \underbrace{D_{KL}(q(x_t \mid x_{t-1,x_0}) \parallel p_\theta(x_{t-1} \mid x_t, y))}_{L_{t-1}} - \underbrace{\log p_\theta(x_0 \mid x_1, y)}_{L_0}] \tag{16}$$

Based on L_{t-1}, the simplified loss function is rewritten as:

$$L_{simple} = E_{t,x_0,\epsilon_t}[\|\epsilon_t - \epsilon_\theta(x_t, y, t)\|^2] \tag{17}$$

In the generation process, the decoding process at step t is changed as follows:

Using a trained neural network model to predict the mean $\mu_\theta(x_t, y, t)$ and variance $\beta_\theta(x_t, y, t)$ of the data distribution at the step t-1, obtaining the data distribution $p_\theta(x_{t-1} \mid x_t)$.

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(\mu_\theta(x_t, y, t), \beta_\theta(x_t, y, t)) \tag{18}$$

Using the reparameterization, we obtain x_{t-1}:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1 - \overline{\alpha}_t}}\epsilon_\theta(x_t, y, t)) + \epsilon\sqrt{\beta(x_t, y, t)}, \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{19}$$

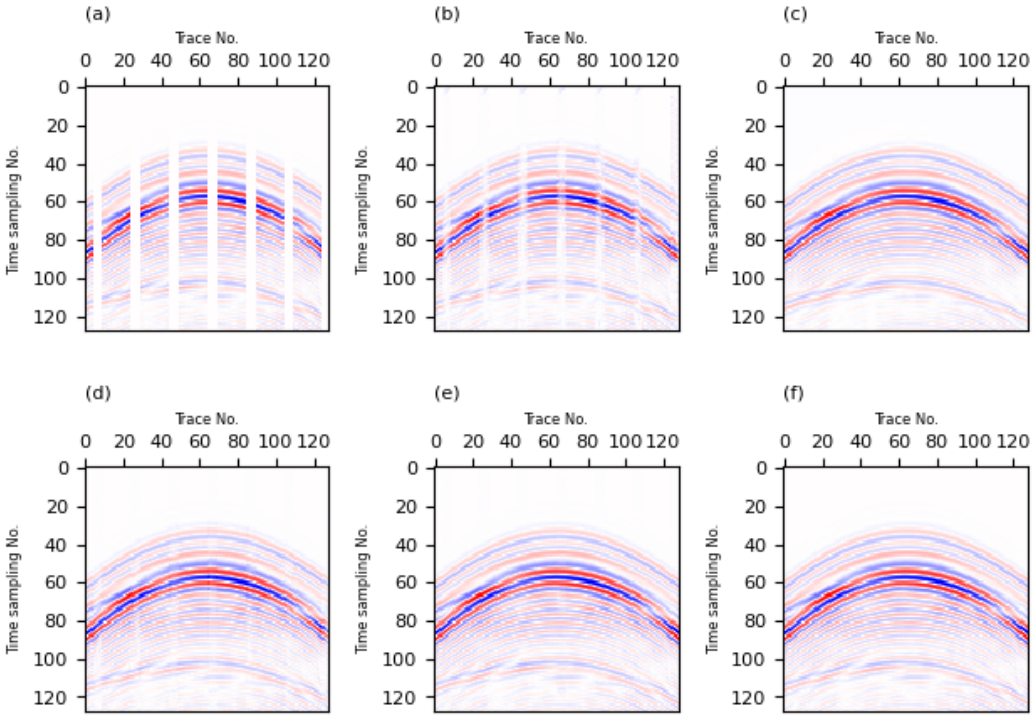

Figure 1. The reconstruction results of different networks for small gap missing traces in the synthetic dataset: (a) Interpolated data, (b) CAE, (c) POCSCNN, (d) ANet, (e) Diffusion model, (f) Ground truth.

4 Experiments

4.1 Synthetic Data

To assess the effectiveness of the proposed method, we conducted experiments on a synthetic dataset using the publicly available Society of Exploration Geophysicists (SEG) C3 dataset. This dataset consists of 45 shots sampled at an 8 ms rate, with each shot containing a receiver grid of size 201 $\times$ 201 and 625 samples per trace.

A total of 1800 patches were selected, out of which 1260 patches were used for training, 360 patches for validation and 180 patches for testing. The value of T for the forward process was set to 1000, and the number of resampling steps was set to 250.

In addition, three different network models were selected for comparative testing, including CAE (Y. Wang et al., 2020), POCSCNN (Zhang et al., 2020), and ANet (Yu & Wu, 2021). Following the methods described in the paper, these models were trained to their optimal states and then compared.

Fig. 1 shows the reconstruction results of the four network models for small gap missing traces. In the data, 28% of the traces were intentionally set to 0 to represent the missing traces, which were distributed in seven locations with each location accounting for 4% of the data. The results demonstrate that CAE did not perform well in the task of reconstruction, POCSCNN exhibited relatively satisfactory results but introduced certain biases, and ANet achieved slightly improved results while still exhibiting some data biases. Conversely, our proposed method yielded the most plausible and reasonable results.

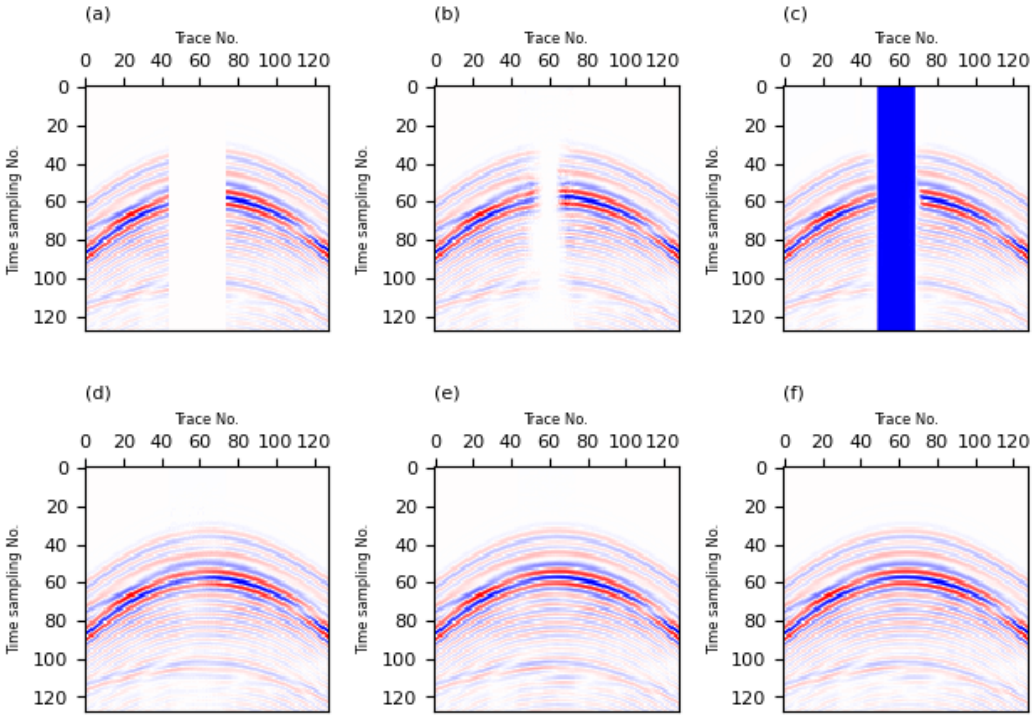

Figure 2. The reconstruction results of different networks for large gap missing traces in the synthetic dataset: (a) Interpolated data, (b) CAE, (c) POCSCNN, (d) ANet, (e) Diffusion model, (f) Ground truth.

To ensure an accurate assessment of the reconstruction results, three commonly employed metrics were employed. Specifically, the Mean Squared Error (MSE), Mean Absolute Error (MAE), and Structural Similarity (SSIM) (Huang et al., 2022) were computed to quantify the disparities between the reconstructed data and the ground truth. The SSIM metric was employed to gauge the resemblance between the two datasets, with values ranging from 0 to 1. A higher SSIM value indicates a greater likeness between the datasets. The comparison of the four network models (see Table S1 in Supporting Information S1) revealing that our proposed method outperforms the other methods in terms of these metrics, demonstrating superior performance.

To evaluate the reconstruction performance of the diffusion model in the context of large gap missing traces, 25% of the consecutive traces in the data were intentionally set to 0 to represent the missing traces. The results were compared with the other three models, as shown in Fig. 2. It can be observed that CAE and POCSCNN performed the worst, with CAE only reconstructing a portion of the traces near the known part, and POCSCNN even experiencing failure in reconstruction. ANet lost some details and had slightly inferior performance compared to the method proposed in this paper. The comparison results of the four networks (see Table S2 in Supporting Information S1) show that our method still exhibited the best performance.

4.2 Field Data

To assess the effectiveness of our method on field data, we conducted experiments on the Mobil Avo Viking Graben Line 12 field dataset and compared it with

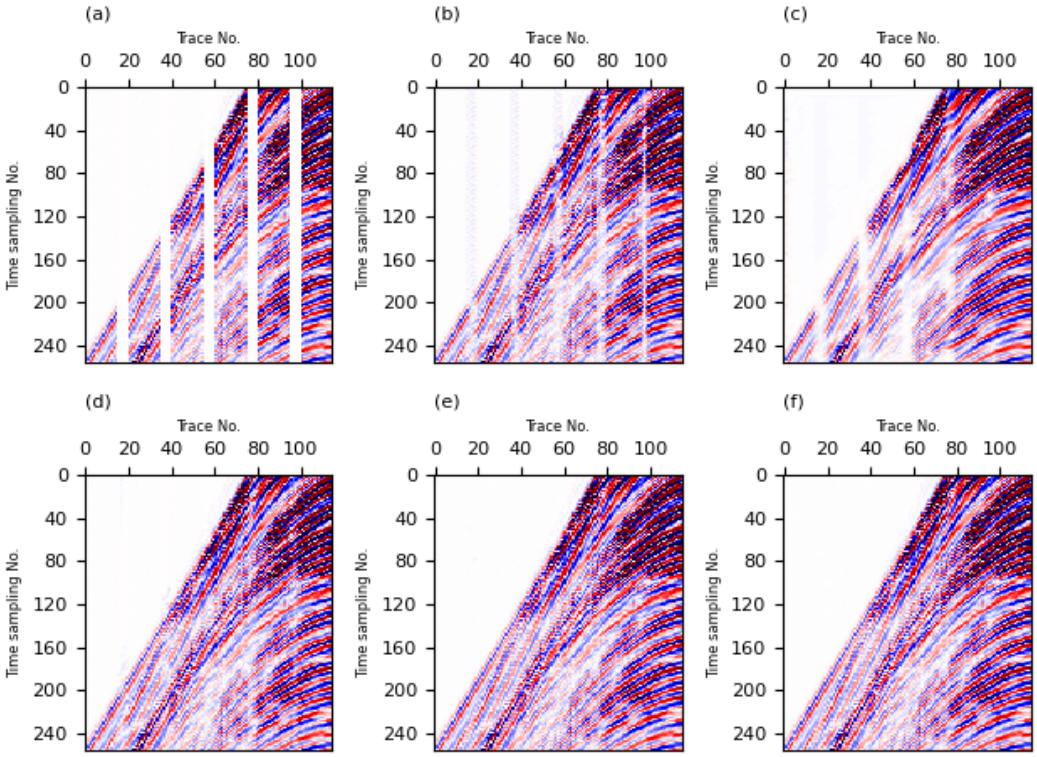

Figure 3. The reconstruction results of different networks for small gap missing traces in the field dataset: (a) Interpolated data, (b) CAE, (c) POCSCNN, (d) ANet, (e) Diffusion model, (f) Ground truth.

three other models. A total of 1000 patches were selected, with 700 patches allocated for training, 200 patches for validation and 100 patches for testing. The value of T for the forward process was set to 1000, and the number of resampling steps was set to 250.

Fig. 3 shows the reconstruction results of the four network models for small gap missing traces in the field dataset. In the data, 20% of the traces were intentionally set to 0 to represent the missing traces, which were distributed in five locations with each location accounting for 4% of the data. It is evident that CAE did not perform well in the reconstruction task, POCSCNN yielded slightly improved results but introduced certain biases, and ANet approached the correct reconstruction but still exhibited some data biases. In contrast, our proposed method produced the most reasonable results.

Calculates the MSE, MAE, and SSIM between the reconstructed data and the ground truth (see Table S3 in Supporting Information S1), it can be observed that our method outperforms the other methods significantly in these metrics, demonstrating superior performance.

To evaluate the reconstruction performance of the diffusion model on large gap missing traces in the field dataset, 25% of the continuous traces in the data were set to 0 as missing traces. A comparison was made with other three models, and the results are shown in Fig. 4. It can be observed that CAE and POCSCNN performed the worst. CAE only reconstructed partial traces near the known part, while POCSCNN even failed to reconstruct. ANet missed some details and had slightly worse performance compared to our method. The comparison of the four networks (see Table S4 in Supporting Information S1), show that our method still exhibited

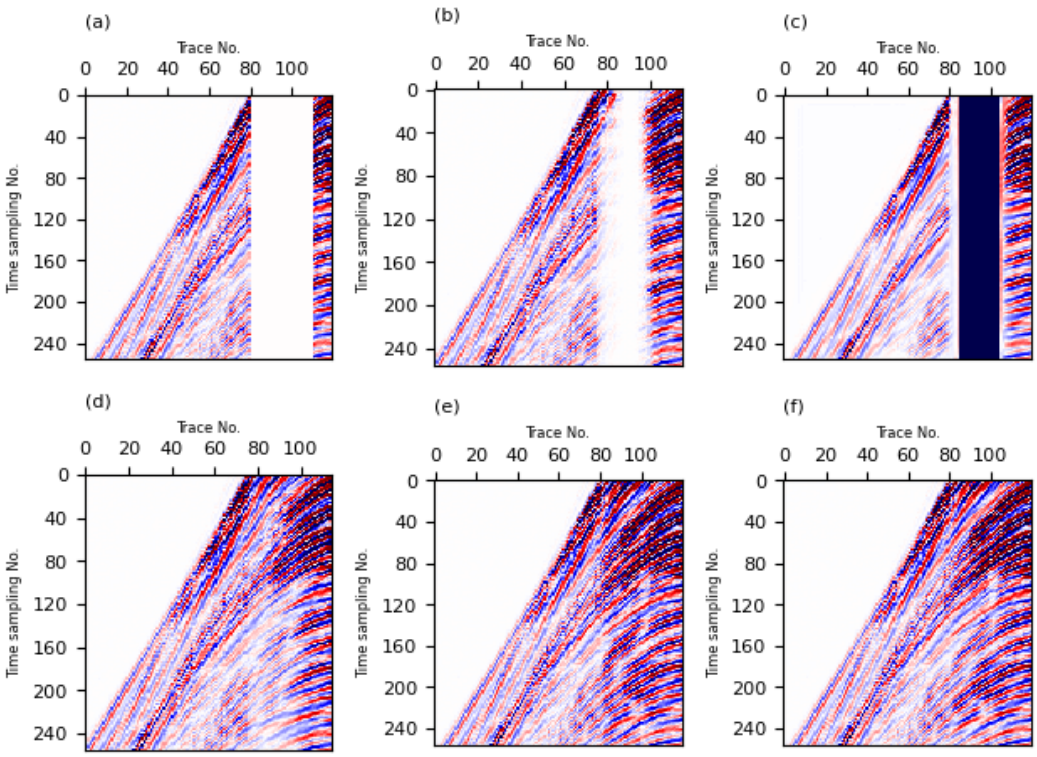

Figure 4. The reconstruction results of different networks for large gap missing traces in the field dataset: (a) Interpolated data, (b) CAE, (c) POCSCNN, (d) ANet, (e) Diffusion model, (f) Ground truth.

superior performance. The above experiments thoroughly validate the effectiveness and applicability of the proposed method in this study.

5 Conclusions

This paper presents a constrained diffusion model for seismic data interpolation and utilize Resampling to impose additional constraints by sampling from given data during the reverse diffusion iterations. This marks the first successful application of the diffusion model in seismic data reconstruction. By learning the distribution of existing seismic data, this method effectively mitigates substantial deviations between generated data and ground truth, which are caused by encoding distortions in traditional convolutional discriminative networks. Comparative experimental results on synthetic and field datasets substantiate that our proposed method achieves more accurate interpolation results compared to existing methods. Additionally, the diffusion model exhibits superior generative accuracy and enhanced generalization ability by learning the data distribution, enabling our network generalizes to more complexity missing scenario during the inference period.

6 Open Research

The observation of Society of Exploration Geophysicists (SEG) C3 dataset (Aminzadeh et al., 1997) of this study are available at `https://wiki.seg.org/wiki/SEG_C3_45_shot`.

The observation of Mobil Avo Viking Graben Line 12 field dataset (Keys & Foster, 1998) of this study are available at `https://wiki.seg.org/wiki/Mobil_AVO_viking_graben_line_12`.

Acknowledgments

The authors sincerely thank Professors Fei Deng for the advice for this study, thanks Peifan Jiang and Bin Wang for the advice.

References

Aminzadeh, F., Brac, J., & Kunz, T. (1997). 3d salt and overthrust models: Presented at the seg. *EAGE Modeling Series*(1).

Chai, X., Gu, H., Li, F., Duan, H., Hu, X., & Lin, K. (2020). Deep learning for irregularly and regularly missing data reconstruction. *Scientific reports*, *10*(1), 3302.

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, *34*, 8780–8794.

He, T., Wu, B., & Zhu, X. (2021). Seismic data consecutively missing trace interpolation based on multistage neural network training process. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, *33*, 6840–6851.

Huang, H., Wang, T., Cheng, J., Xiong, Y., Wang, C., & Geng, J. (2022). Self-supervised deep learning to reconstruct seismic data with consecutively missing traces. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–14.

Jia, Y., & Ma, J. (2017). What can machine learning do for seismic data processing? an interpolation application. *Geophysics*, *82*(3), V163–V177.

Keys, R. G., & Foster, D. J. (1998). A data set for evaluating and comparing seismic inversion methods. *Comparison of seismic inversion methods on a single real data set*, 1–12.

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

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

Kuijpers, D., Vasconcelos, I., & Putzky, P. (2021). Reconstructing missing seismic data using deep learning. *arXiv preprint arXiv:2101.09554*.

Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., & Van Gool, L. (2022). Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 11461–11471).

Naghizadeh, M., & Sacchi, M. (2009). fx adaptive seismic-trace interpolation: Geophysics, 74. *V9–V16*.

Nichol, A. Q., & Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *International conference on machine learning* (pp. 8162–8171).

Pan, S., Chen, K., Chen, J., Qin, Z., Cui, Q., & Li, J. (2020). A partial convolution-based deep-learning network for seismic data regularization1. *Computers & Geosciences*, *145*, 104609.

Park, J., Choi, J., Jee Seol, S., Byun, J., & Kim, Y. (2021). A method for adequate selection of training data sets to reconstruct seismic data using a convolutional u-net. *Geophysics*, *86*(5), V375–V388.

Pawelec, I., Wakin, M., & Sava, P. (2021). Missing trace reconstruction for 2d land seismic data with randomized sparse sampling. *Geophysics*, *86*(3), P25–P36.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the*

ieee/cvf conference on computer vision and pattern recognition (pp. 10684–10695).

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

Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y., & Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, *32*.

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

Wang, Y., Wang, B., Tu, N., & Geng, J. (2020). Seismic trace interpolation for irregularly spatial sampled data using convolutional autoencodercae-based seismic trace interpolation. *Geophysics*, *85*(2), V119–V130.

Wei, Q., Li, X., & Song, M. (2021). Reconstruction of irregular missing seismic data using conditional generative adversarial networks. *Geophysics*, *86*(6), V471–V488.

Yang, P., Gao, J., & Chen, W. (2012). Curvelet-based pocs interpolation of nonuniformly sampled seismic records. *Journal of Applied Geophysics*, *79*, 90–99.

Yu, J., & Wu, B. (2021). Attention and hybrid loss guided deep learning for consecutively missing seismic data reconstruction. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–8.

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

Zhou, Y., & Han, C. (2018). Seismic data restoration based on the grassmannian rank-one update subspace estimation method. *Journal of Applied Geophysics*, *159*, 731–741.

Supporting Information for "A Research Of Seismic Data Reconstruction Based On Conditional Constraint Diffusion Model"

Fei Deng[1], Shuang Wang[2], Xuben Wang[3]and Peng Fang[4]

[1]College of Computer Science and Cyber Security, Chengdu University of Technology, Chengdu, China

[2]College of Computer Science and Cyber Security, Chengdu University of Technology, Chengdu, China

[3]College of Geophysics,Chengdu University of Technology, Chengdu, China

[4]Chinese academy of science, Institute of geology and geophysics, Peking, China

Contents of this file

Tables S1 to S4

Introduction

Table S1 calculates the MSE, MAE, and SSIM between the reconstructed data and the ground truth for the four networks on synthetic dataset with small gap missing traces.

Table S2 calculates the MSE, MAE, and SSIM between the reconstructed data and the ground truth for the four networks on synthetic dataset with large gap missing traces.

Table S3 calculates the MSE, MAE, and SSIM between the reconstructed data and the ground truth for the four networks on field dataset with small gap missing traces.

Table S4 calculates the MSE, MAE, and SSIM between the reconstructed data and the ground truth for the four networks on field dataset with large gap missing traces.

Table S1. comparison of four reconstruction networks for small gap missing traces in the synthetic dataset.

	MSE	MAE	SSIM
CAE	1.9647	0.21	0.80
POCSCNN	0.1901	0.13	0.62
ANET	0.4251	0.09	0.89
Diffusion Model	0.0384	0.03	0.94

Table S2. Comparison of four reconstruction networks for large gap missing traces in the synthetic dataset.

	MSE	MAE	SSIM
CAE	2.7150	0.22	0.82
POCSCNN	100.1771	2.05	0.53
ANET	0.7328	0.13	0.89
Diffusion Model	0.1388	0.06	0.93

Table S3. Comparison of four reconstruction networks for small gap missing traces in the field dataset.

	MSE	MAE	SSIM
CAE	55.0735	1.34	0.82
POCSCNN	48.8067	2.35	0.50
ANET	19.7354	0.84	0.86
Diffusion Model	0.8422	0.18	0.93

Table S4. Comparison of four reconstruction networks for large gap missing traces in the field dataset.

	MSE	MAE	SSIM
CAE	227.0666	3.61	0.78
POCSCNN	7882.4298	26.52	0.33
ANET	148.1965	2.90	0.82
Diffusion Model	28.0099	1.34	0.92