https://doi.org/10.5194/essd-2024-337
Preprint. Discussion started: 7 October 2024

Open Access Earth System Science Data Discussions

cigFacies: a massive-scale benchmark dataset of seismic facies and its application

Hui Gao[1], Xinming Wu[1], Xiaoming Sun[2], Mingcai Hou[3,4], Hang Gao[1], Guangyu Wang[1], and Hanlin Sheng[1]

[1]School of Earth and Space Sciences, University of Science and Technology of China, Hefei, China;
[2]Institute of Advanced Technology, University of Science and Technology of China, Hefei, China;
[3]Institute of Sedimentary Geology, Chengdu University of Technology, Chengdu, China;
[4]State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation, Chengdu University of Technology, Chengdu, China;

Correspondence: Xinming Wu (xinmwu@ustc.edu.cn)

Abstract. Seismic facies classification is crucial for seismic stratigraphic interpretation and hydrocarbon reservoir characterization but remains a tedious and time-consuming task that requires significant manual effort. The data-driven deep learning approaches are highly promising to automate the seismic facies classification with high efficiency and accuracy, as they have already achieved significant success in similar image classification tasks within the field of computer vision (CV). However, unlike the CV domain, the field of seismic exploration lacks a comprehensive benchmark dataset for seismic facies, severely limiting the development, application, and evaluation of deep learning approaches in seismic facies classification. To address this gap, we propose a comprehensive workflow to construct a massive-scale benchmark dataset of seismic facies and evaluate its effectiveness in training a deep learning model. Specifically, we first develop a knowledge graph of seismic facies based on the geological concepts and seismic reflection configurations. Guided by the graph, we then implement three strategies of field seismic data curation, knowledge-guided synthesization, and GAN-based generation to construct a benchmark dataset of 8000 diverse samples for five common seismic facies. Finally, we use the benchmark dataset to train a network and then apply it on two 3-D seismic data for automatic seismic facies classification. The predictions are highly consistent with expert interpretation results, demonstrating the diversity and representativeness of our benchmark dataset is sufficient to train a network that can generalize well in seismic facies classification across field data. We have made this dataset, the trained model and associated codes publicly available for further research and validation of intelligent seismic facies classification.

1 Introduction

Seismic facies classification aims to delineate individual units based on the specific reflection characteristics (e.g. reflection configuration, continuity, amplitude and frequency contents), which is a fundamental and essential step in the seismic stratigraphic analysis and contributes to the interpretation of sedimentary environments and hydrocarbon reservoir distributions (Sheriff, 1976; Sangree and Widmier, 1977; Veeken, 2006; Jia and Zhao, 2007; Xu and Haq, 2022). With the dramatic increase in the amount of 3-D seismic data, manual interpretation method is typically labor-intensive and heavily relies on the

https://doi.org/10.5194/essd-2024-337
Preprint. Discussion started: 7 October 2024

experienced experts, thus the automatic seismic facies classification is the trend. Moreover, the development of automatic seismic facies classification approaches benefits the accurate and efficient analysis of depositional environments and lithologic distributions.

In recent years, many methods have been proposed for automatic seismic facies classification by using supervised, semi-supervised and unsupervised learning. Supervised learning methods (Wrona et al., 2018; Zhao, 2018; Liu et al., 2018; Zhang et al., 2021) first use large amounts of labeled data to train a CNN model, and then use the trained model for automatic seismic facies classification. Semi-supervised learning methods (Qi et al., 2016; Dunham et al., 2020; Liu et al., 2020) use both labeled and unlabeled data to train the network to learn the features and distributions characterizing seismic facies. Unsupervised learning methods (Qian et al., 2018; Zhao et al., 2018; Duan et al., 2019; Puzyrev and Elders, 2022; Li et al., 2023) first extract the nonlinear, discriminant and invariant features from unlabeled data, and then cluster or classify these features for automatic seismic facies classification. The supervised learning methods often exhibit weak generalization capabilities across different surveys due to a lack of labeled samples, while semi-supervised and unsupervised methods frequently encounter issues with high uncertainty in prediction results. Besides, seismic facies can be classified into several different categories based on different attribute parameters, which leads to challenges in the construction of seismic facies datasets and the assessment of the results.

To solve these problems, developing a knowledge graph of seismic facies and using it to provide guidelines for constructing a benchmark dataset is considered an effective methodology. Knowledge graph is a graphical representation model consisting of entities (nodes) and relationships (edges), which aims to represent knowledge in the form of graphs (Paulheim, 2017; Fensel et al., 2020; Hogan et al., 2021). Currently, knowledge-driven geoscience big data researches have been successfully applied in various kinds of geoscience data-mining tasks (Zhou et al., 2021; Ma et al., 2023; Zhang et al., 2023; Hu et al., 2023). In this work, we construct a knowledge graph of seismic facies, grounded in geological concepts and seismic reflection patterns. This graph guides our processes of data selection, label generation, analysis, and result assessment.

To address the lack of representative benchmark datasets for seismic facies and improve its automatic classification, we introduce a workflow (Fig. 1) to construct a massive-scale, feature-rich and high-realism benchmark dataset of seismic facies and use it to train a CNN model for accurate and efficient seismic facies classification. Initially, we construct a knowledge graph of seismic facies based on the geological concepts and seismic facies configurations. Guided by the graph, we develop three strategies of field data curation, knowledge-guided synthesization and Generative Adversarial Network (GAN)-based generation to construct a massive-scale benchmark dataset of seismic facies. Utilizing this diverse dataset, we train a CNN model and subsequently apply it to two 3-D field seismic datasets for automatic seismic facies classification.

2 Building a massive-scale benchmark dataset of seismic facies

In this section, we implement a workflow of three strategies shown in left blue box in Fig. 1 to construct a massive-scale, feature-rich and high-realism benchmark dataset of seismic facies. The first strategy is to build field samples from field data curation with raw data collection, data standardization and skeletonization processes. The second strategy is to build synthetic

https://doi.org/10.5194/essd-2024-337
Preprint. Discussion started: 7 October 2024

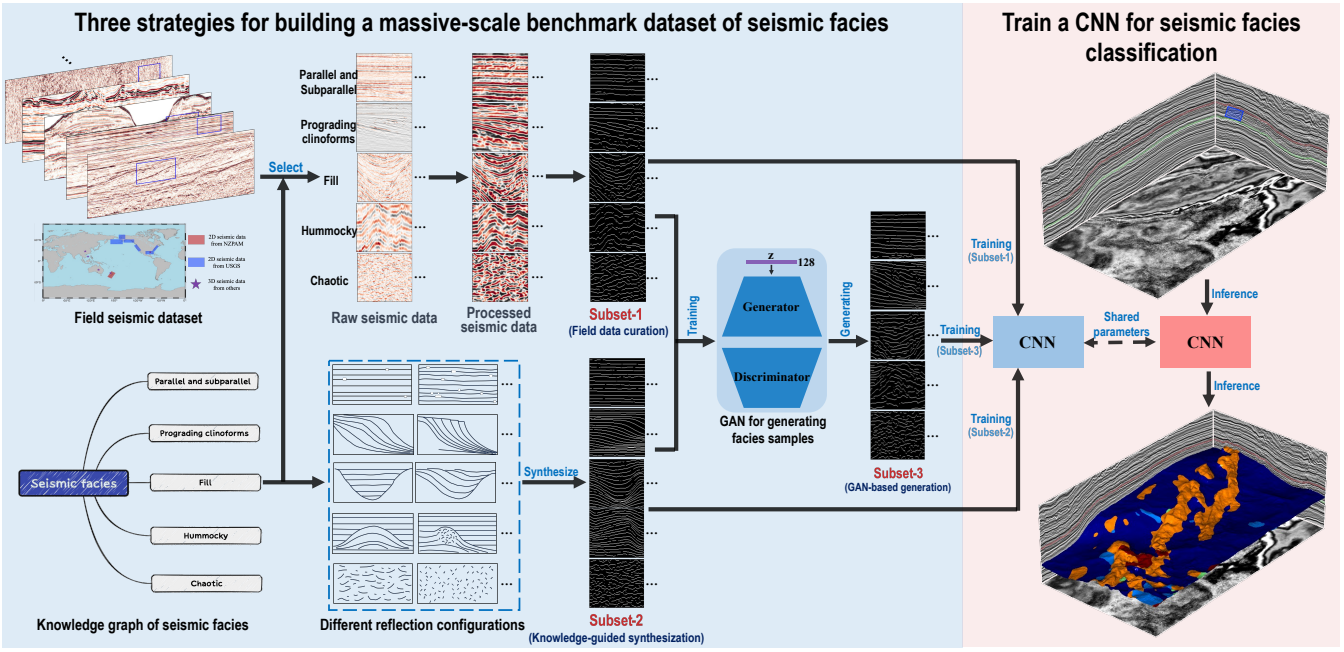

Figure 1. The workflow of constructing a massive-scale, feature-rich and high-realism benchmark dataset of seismic facies (blue box) and deep learning for seismic facies classification (red box). We first develop a knowledge graph of seismic facies based on geological concepts and seismic facies configurations. Guided by the graph, we implement three strategies of field seismic data curation, knowledge-guided synthesization, and AI-based generation to construct a massive-scale benchmark dataset. Finally, we use the benchmark dataset to train a CNN model and then apply it on 3-D field seismic data for automatic seismic facies classification.

samples from knowledge-guided synthesization by synthesizing geological structural curves. The final strategy is to build synthetic samples from AI-based generation with GAN model. By employing these three strategies, we have ultimately constructed a dataset containing 2000, 1500, 1500, 1500, and 1500 samples for five common seismic facies, respectively.

2.1 Knowledge graph of seismic facies

Before constructing the massive-scale benchmark dataset of seismic facies, it is necessary to develop a knowledge graph of seismic facies based on the geological concepts and seismic reflection configurations, which can provide guidelines for preparing representative dataset samples and assessing facies classification results. Based on specific seismic reflection configurations, seismic facies can be roughly divided into parallel and subparallel, prograding clinoforms, fill, hummocky, chaotic, divergent, wave and reflection free (Mitchum Jr et al., 1977a, b; Veeken, 2006; Xu and Haq, 2022) (Fig. 2). Besides, these seismic facies can be further subdivided based on several independent parameters such as the reflection configurations, continuity, amplitude and frequency. For example, parallel and subparallel reflection can be subdivided into 27 different types based on the frequency (high, middle and low), amplitude (strong, moderate and week) and continuity (excellent, medium and poor). Based on different reflection patterns, prograding clinoforms, fill and hummocky can be futher subdivided into five (sigmoid, oblique,

https://doi.org/10.5194/essd-2024-337
Preprint. Discussion started: 7 October 2024

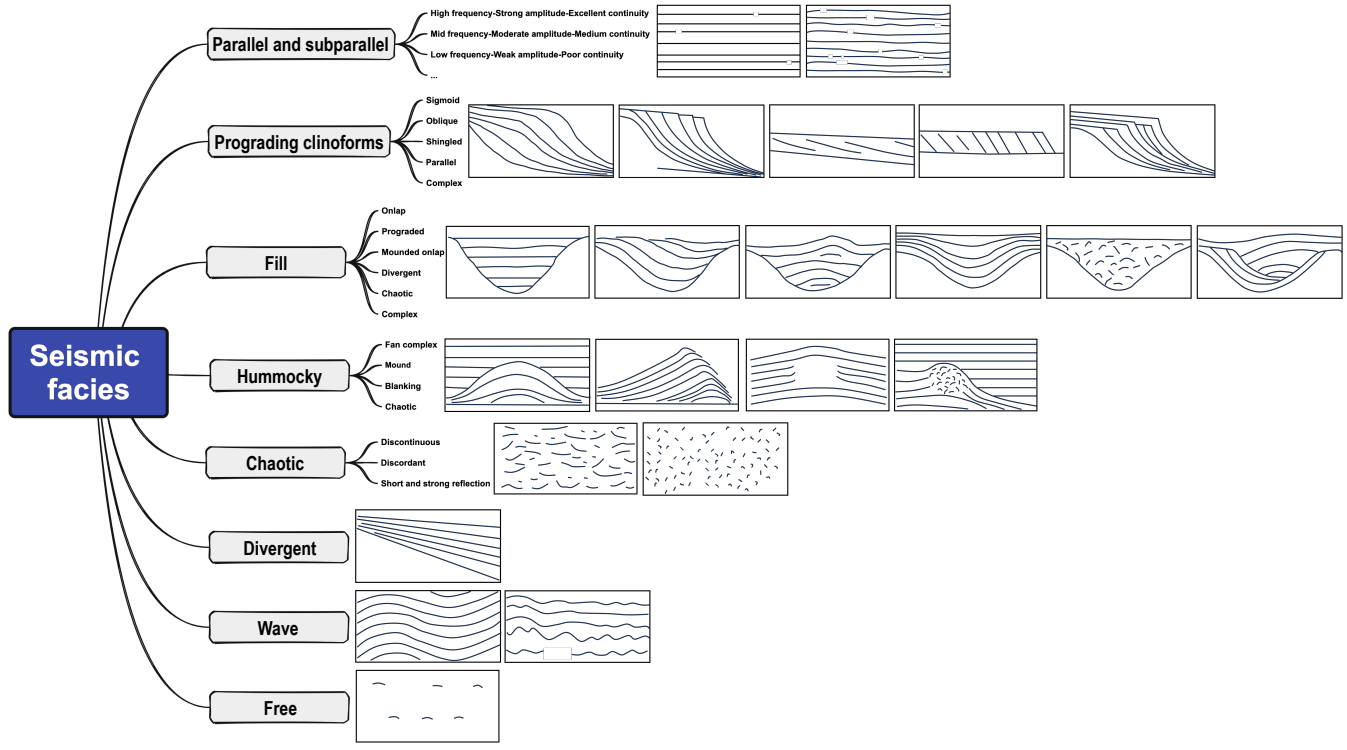

Figure 2. Knowledge graph of seismic facies and corresponding typical seismic reflection configurations (modified from Mitchum Jr et al. (1977a, b); Xu and Haq (2022)). In this graph, we roughly divided the seismic facies into eight types (parallel and subparallel, prograding clinoforms, fill, hummocky, chaotic, divergent, wave and reflection free) based on specific seismic reflection configurations. Besides, we also subdivided these seismic facies based on several independent parameters such as the reflection configurations, continuity, amplitude and frequency, and illustrate the typical seismic reflection configurations for each type of seismic facies.

shingled, parallel and complex), six (onlap, prograded, mounded onlap, divergent, chaotic and complex) and four (fan complex, mound, blanking and chaotic) types, respectively.

As shown in Fig. 2, we develop a knowledge graph of seismic facies and illustrate the typical seismic reflection configurations for eight types of seismic facies. However, considering the requirement for data amount and diversity in this work, we take the five most common seismic facies (parallel and subparallel, prograding clinoforms, fill, hummocky, and chaotic) as an example to explain how to construct a massive-scale, feature-rich and high-realism benchmark dataset of seismic facies from field data curation, knowledge-guided synthesization and GAN-based generation.

2.2 Building facies samples by field data curation

We start building our benchmark dataset by employing the field seismic data curation strategy with a series of steps including raw data collection, manual interpretation and classification, bandpass filtering, resampling, amplitude equalization, and

https://doi.org/10.5194/essd-2024-337
Preprint. Discussion started: 7 October 2024

skeletonization. We first collect almost 4000 global publicly available 2-D seismic profiles and 10 3-D seismic data from the sources of United States Geological Survey (USGS), New Zealand Petroleum And Minerals (NZPAM), South Australian Resources Information Gateway (SARIG), Society of Exploration Geophysicists (SEG) and so on. These 2-D and 3-D seismic data amount to around 130G, primarily located in the Gulf of Mexico, East and West Coast of America, Alaska, Bering Sea, Beaufort Sea, New Zealand, Southern Australia and Sichuan Basin (see the data distributions map in Figure 1).

We then manually select, crop and classify these field seismic data based on the knowledge graph (Fig. 2). As shown in the raw seismic data of Fig. 3, we totally collect 1000, 700, 500, 500 and 700 2-D raw seismic data for five common seismic facies, respectively. However, due to the different data sources, depositional environments and data processing methods, these raw seismic data have large differences in sampling rates, amplitude and frequency distributions (as shown in Fig. 3 and Fig. 4a) among same and different classes of seismic facies. These data variations and uncertainties are not related to the seismic facies. Moreover, they may pose significant inference to deep learning models in learning the crucial features such as texture patterns and reflection configurations, which are essential for identifying seismic facies categories. To migrate such uncertainties in building our standard benchmark dataset, we introduce the data standardization process (Fig. 4) for each raw seismic data, including filtering, resampling, amplitude equalization, frequency equalization and so on. After applying the data standardization process, the processed seismic data have been significantly improved in the consistency of the sampling rates, amplitude and frequency distributions (as shown in Fig. 3b and Fig. 5). Finally, we retain the main geological structure informations of strata by keeping only the waveform peaks as ones and setting elsewhere zeros to obtain the corresponding field skeletonization images shown in Fig. 3c and Fig. 6.

Compared to the skeletonization images (Fig. 4d) obtained directly from raw seismic data, the ones (Fig. 4c) with data standardization can more clearly reflect the geological structure characteristics and enhance the consistency among the same and different classes of seismic facies. The whole curation strategy, perticularly the data standardization processes and skeletonization, eliminates uncertainties inherent in field data from various surveys. This approach retains only the texture patterns associated with seismic facies to produce standardized images for constructing the benchmark dataset. The same processing techniques will also be applied to inference data to ensure that a deep learning model trained on this dataset achieves consistent predictions.

The facies samples from only the field seismic datasets are imbalanced in categaries and lack diversity and therefore are not sufficient to build a massive-scale and representative benchmark dataset. For example, parallel and subparallel data are more common than fill or hummocky data in field seismic data. Additionally, some specific patterns (e.g., parallel prograding clinoforms, chaotic fill, complex fill and blanking hummocky) are rare in these publicly available field seismic data.

2.3 Building facies samples from knowledge-guided synthesization

In order to overcome the sample imbalance and improve the diversity of the dataset, we further develop the second strategy to automatically generate synthetic facies samples based on the knowledge graph of seismic facies and independent seismic

https://doi.org/10.5194/essd-2024-337
Preprint. Discussion started: 7 October 2024

Open Access Earth System Science Data Discussions

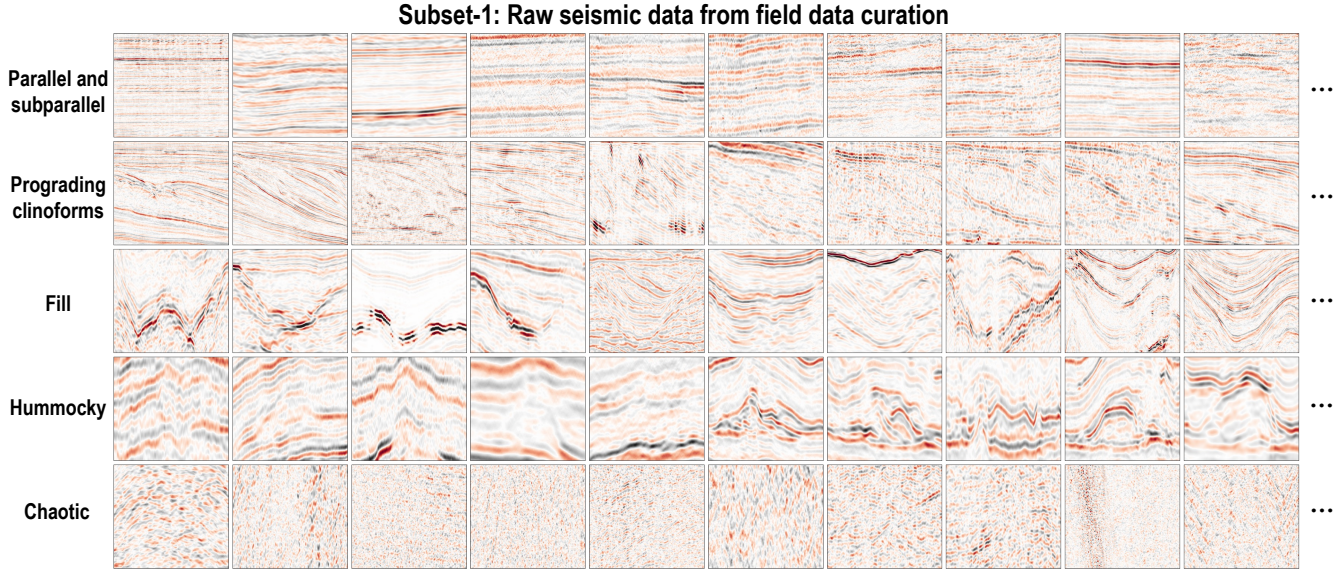

Figure 3. Subset-1: raw seismic data manually collected and interpreted from the large amount of publicly available seismic datasets. In total, we select, crop and classify 1000, 700, 500, 500 and 700 2-D raw seismic data for parallel and subparallel, prograding clinoforms, fill, hummocky and chaotic, respectively.

reflection configurations. We first define different geological structural curves by using following geometric functions:

$$z = z_0, \tag{1}$$

$$z = k_0 \cdot x + z_0, \tag{2}$$

$$z = k_0 \cdot x^2 + z_0, \tag{3}$$

$$z = \frac{1}{k_1 + k_2 \cdot e^{-k_3 \cdot x}}, \tag{4}$$

$$z = \frac{e^{k_1 \cdot x} - e^{-k_2 \cdot x}}{e^{k_1 \cdot x} + e^{-k_2 \cdot x}}, \tag{5}$$

where x and z represent the position in the crossline and depth directions, respectively. Other parameters (z_0, k_0, k_1, k_2 and k_3) are used to control the geometry and distribution of the geological structural curves. Then we randomly combine these geological structural curves at random intervals by using these functions. Furthermore, we can also first define some key points for some complex geological structures, and then generate the corresponding geological structural curves by applying interpolation process. After generating these different geological structural curves, we add random noise to each curve and randomly mask in the local areas to improve the realism of the synthetic data. Finally, we set ones on the geological structural curves and zeros elsewhere to generate the corresponding synthetic skeletonization data.

In this way, we randomly generate synthetic facies samples for each types of seismic facies, especially some specific patterns which are rare in field data curation, thus complementing the benchmark dataset of seismic facies. Finally, we automatically

https://doi.org/10.5194/essd-2024-337
Preprint. Discussion started: 7 October 2024

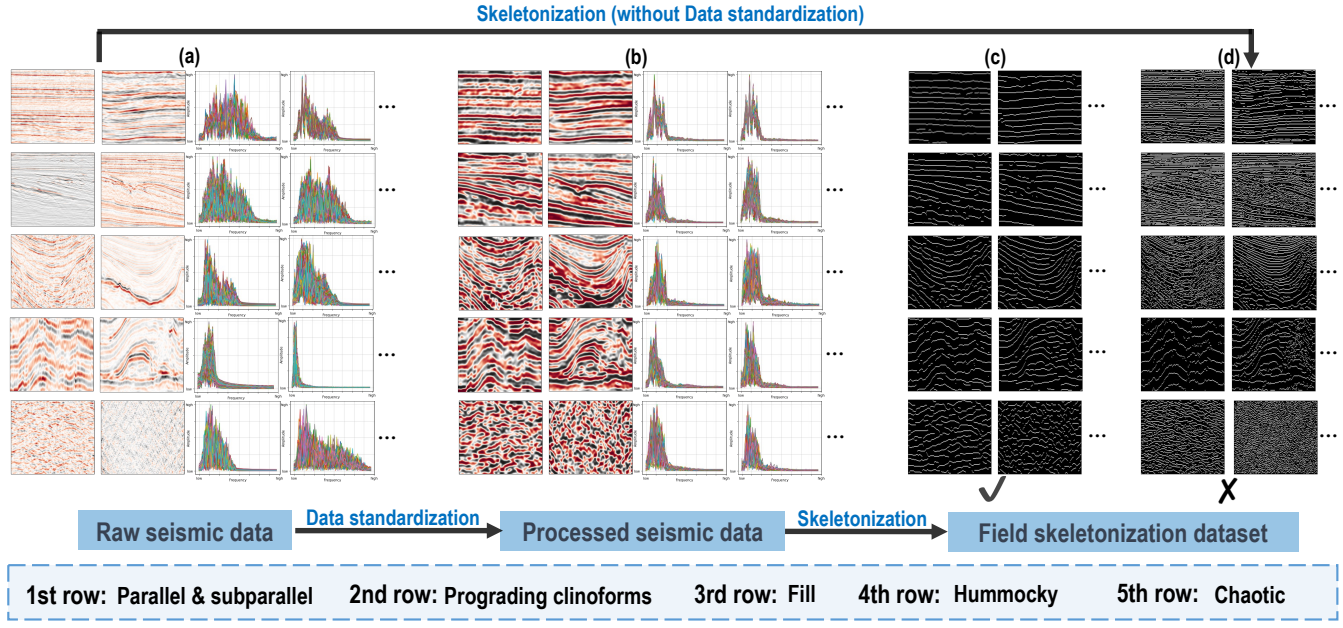

Figure 4. The workflow of constructing the field samples from field seismic data curation. We first manually collect and interpret raw seismic data (a). Then we introduce data standardization process for each raw seismic data to improve the consistency of the sampling rates, amplitude and frequency distributions. After obtaining the processed seismic data (b) , we retain the main geological structure informations of strata by keeping on the waveform peaks as ones and setting elsewhere zeros to obtain the corresponding field skeletonization images (c).

generate 500 synthetic facies samples for five common seismic facies shown in Fig. 7, respectively. Compared to the field facies samples shown in Fig. 6, the synthetic facies samples generated from knowledge-guided synthesization contain more diverse patterns and reduces sample imbalance. However, these synthetic facies samples may be ideally patterned and lack realism.

2.4 Building facies samples from GAN-based generation

As shown in the subset-1 and subset-2 in Fig. 6, Fig. 7 and Fig. 8a, b, the field facies samples has high realism but poor diversity, while the synthetic samples has strong diversity but poor realism. In order to construct a massive-scale, feature-rich and high-realism benchmark dataset of seismic facies, we develop the third strategy of GAN-based generation (Fig. 8) to build more facies samples with high diversity and strong realism.

As shown in Fig. 8c, the architecture of deep neural network used in this work is modified from the progressive growing of GANs proposed by Karras et al. (2017). Traditionally, the progressive growing of GANs consists of a generator model (G) and a discriminator model (D), where G was used to capture the data distribution and generate fake images to resemble the training dataset (real images), and D was used to assess the probability that images are real or fake. The G is composed of a Gen-1 module, five Gen-2 modules and a $\text{Conv}_{1\times1}$ layer, where Gen-1 module consists of a 4×4 convolutional layer and a 3×3 convolutional layer, and Gen-2 module consists of an upsampling layer and two 3×3 convolutional layers. The D is composed

https://doi.org/10.5194/essd-2024-337
Preprint. Discussion started: 7 October 2024

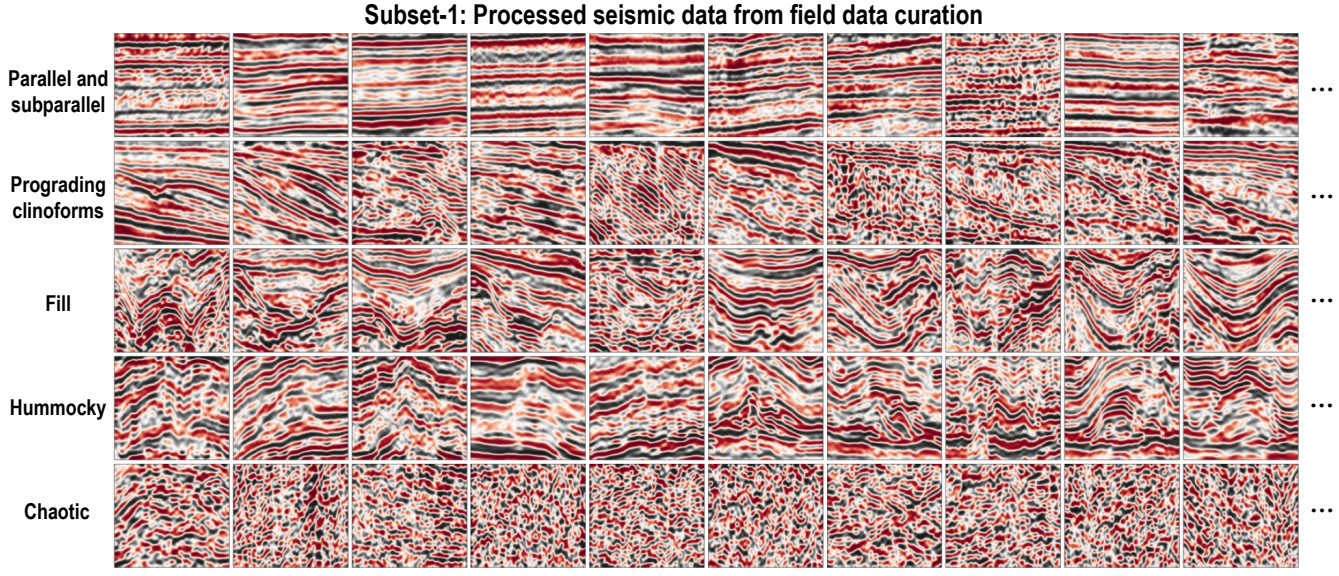

Figure 5. Subset-1: processed seismic data generated from raw seismic data by applying the data standardization processes. Compared to the raw seismic data, the processed seismic data exhibit significant improvement in the consistency of the sampling rates, amplitude and frequency distributions.

of a $\text{Conv}_{1\times1}$ layer, five Dis-1 modules and a Dis-2 module, where Dis-1 module consists of two 3×3 convolutional layer and a average pooling layer, and Dis-2 consists of a minibatch stddev layer, a 3×3 convolutional layer, a 4×4 convolutional layer, a flatten layer and a linear layer. Compared to traditional GANs, the progressive growing of GANs does not directly generate high-resolution images, but starts from generating simple low-resolution images and then continuously increases the resolution of the generated images during the network training. This training strategy allows the network to learn the features of the training dataset from coarse to fine scales, resulting in faster training speed, higher stability and better quality images. Besides, we use WGAN-GP loss proposed by Gulrajani et al. (2017) as the GANs loss function $\mathcal{L}$(G, D) to optimize the network.

We first use subset-1 and subset-2 as training datasets to train the progressive growing of GANs. Initially, we first train a simple network consisting of a Gen-1 module, two $\text{Conv}_{1\times1}$ layer and a Dis-2 module to generate and access the real and fake facies samples with 4×4 scale. After stabilizing the training of this simple network, we then incorporate a Gen-2 module and a Dis-1 module into it for doubling the resolution of G and D. In this way, our network will progressively grow to steadily generate high resolution (128×128) facies samples. Finally, we use the trained G to automatically generate 500 facies samples for each type of seismic facies shown in subset-3 in Fig. 8d and Fig. 9. Compared to the subset-1 and subset-2, the facies samples constructed by the GAN-based generation hold both high diversity and strong realism.

https://doi.org/10.5194/essd-2024-337
Preprint. Discussion started: 7 October 2024

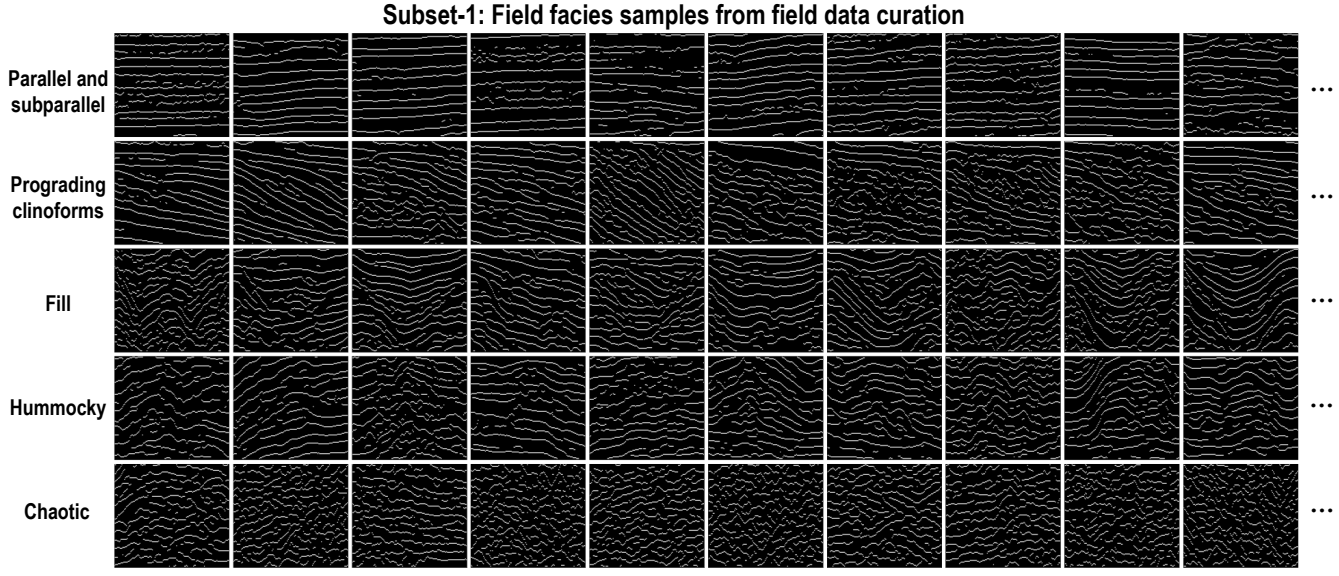

Figure 6. Subset-1: field facies samples generated from processed seismic data by applying the skeletonization process. After obtaining the processed seismic data, we retain the main geological structure informations of strata to generate the corresponding field skeletonization images. Finally, we use the first strategy to manually select 1000, 700, 500, 500 and 700 field facies samples for five common seismic facies, respectively.

2.5 The final benchmark dataset of seismic facies

After applying three strategies of field data curation, knowledge-guided synthesization and GAN-based generation to build diverse facies samples, we construct a massive-scale, feature-rich and high-realism benchmark dataset of seismic facies and we display some facies samples in Fig. 6-9. As shown in Fig. 10, we finally generate a total of 2000, 1500, 1500, 1500, and 1500 diverse facies samples for five common seismic facies (parallel and subparallel, prograding clinoforms, fill, hummocky and chaotic), respectively. The final benchmark dataset, named cigFacies, has been made publicly available at https://zenodo.org/records/10777460 (Gao et al., 2024a).

3 Deep learning for seismic facies classification

After constructing the comprehensive benchmark dataset of seismic facies (Fig. 10), we use it to train a simple CNN for the seismic facies classification task shown in right red box in Fig. 1. In this study, we first use it to train and validate a simple CNN model with 6400 and 1600 pairs of facies samples, respectively. Then we develop a predicted workflow to apply the trained network for automatic seismic facies classification in the 3-D field seismic data.

https://doi.org/10.5194/essd-2024-337
Preprint. Discussion started: 7 October 2024

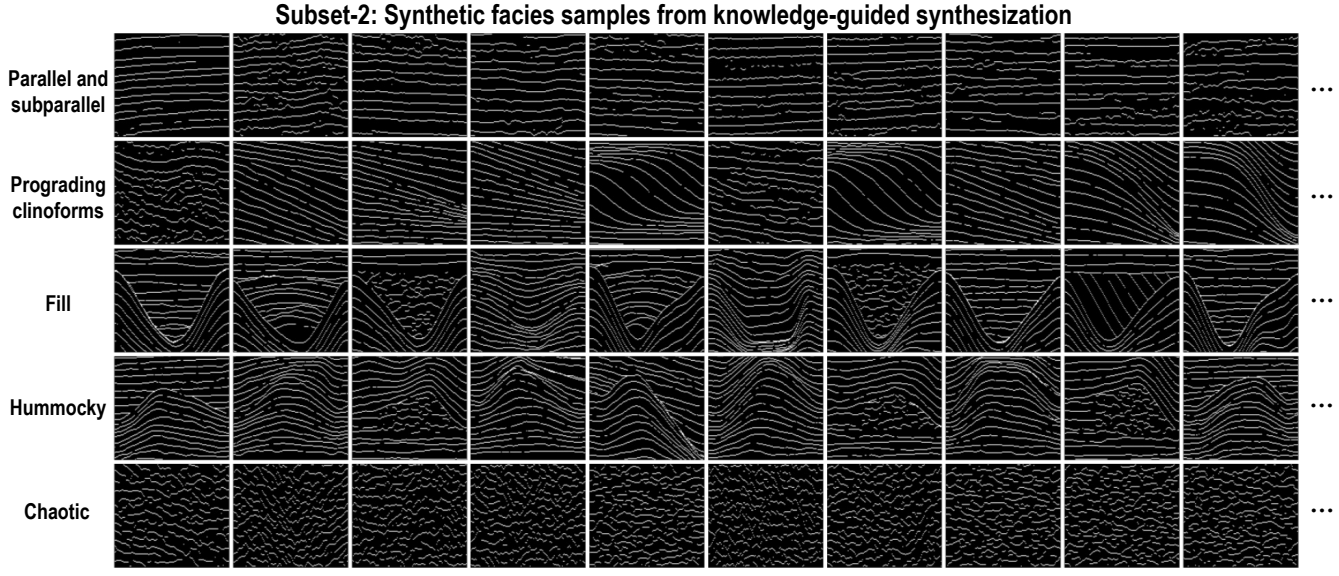

Figure 7. Subset-2: synthetic facies samples generated from knowledge-guided synthesization. In this strategy, we first construct some geological structural curves from geometric functions or interpolation process. Then we add random noise and mask for each curve to improve the realism of synthetic facies samples. Finally, we use the second strategy to automatically generate 500 synthetic facies samples with more diverse patterns for each seismic facies, respectively.

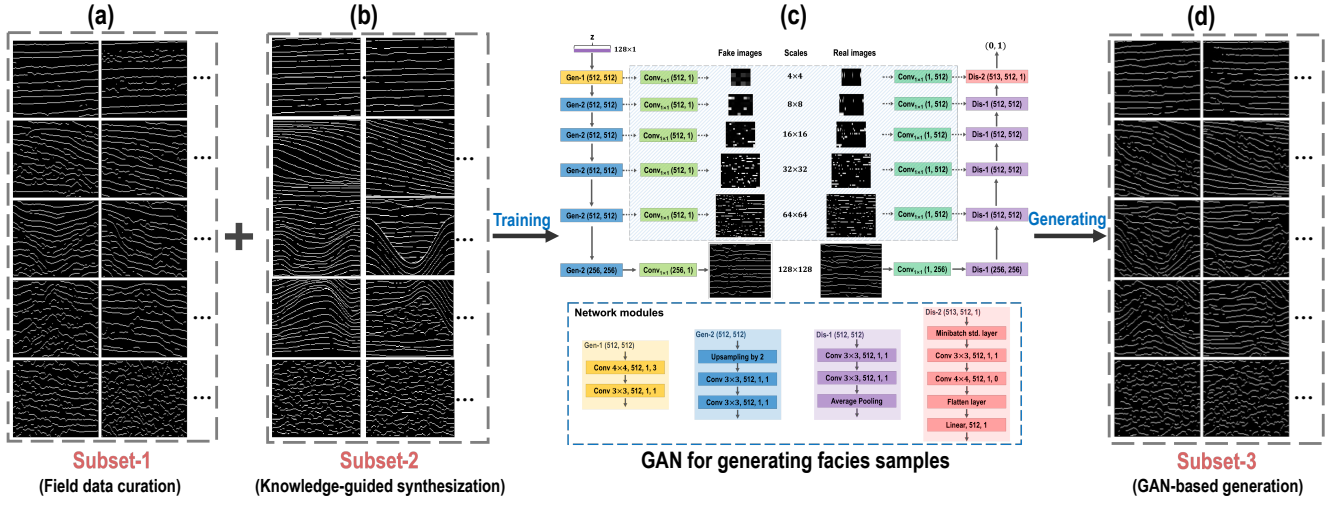

Figure 8. The workflow of constructing the synthetic samples from GAN-based generation. In this strategy, we first use the subset-1 (a) and subset-2 (b) generated from the first and second strategies to train a progressive growing of GANs (c), and then use the trained G to automatically generate synthetic facies samples (d) for each type of seismic facies.

https://doi.org/10.5194/essd-2024-337
Preprint. Discussion started: 7 October 2024

Open Access Earth System Science Data Discussions

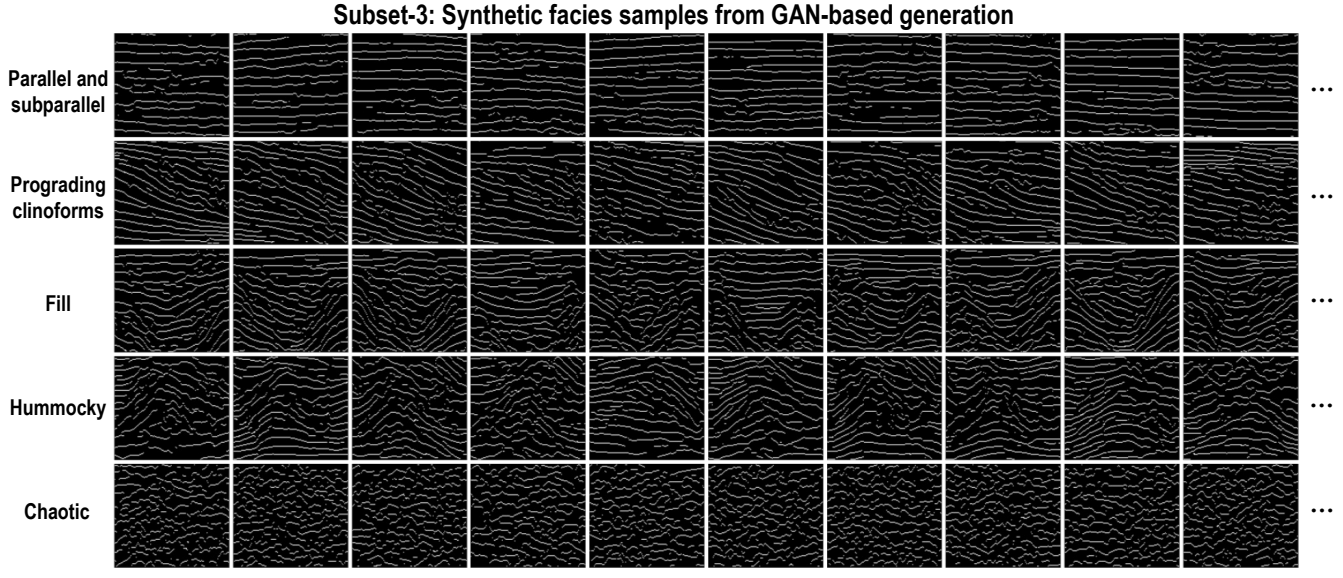

Figure 9. Subset-3: synthetic facies samples generated from GAN-based generation. In total, we use the third strategy to automatically generate 500 synthetic facies samples with both high diversity and strong realism for each type of seismic facies, respectively.

3.1 Training and Validation

We consider seismic facies classification as an image classification problem with the goal to classify the 3-D field seismic data to the corresponding seismic facies (e.g. parallel and subparallel, prograding clinoforms, fill, hummocky and chaotic). In this study, we use a simple deep neural network (ResNet-50) proposed by He et al. (2016) (Fig. 11a) to implement automatic seismic facies classification. We train and validate our CNN model by using 6400 and 1600 random pairs of the benchmark dataset of seismic facies. Besides, in order to improve the diversity of dataset, we apply random data augmentation strategies (e.g. flip, translation, crop and resize) for each facies sample before feeding it into the network. we train our network by using the following cross entropy loss function $\mathcal{L}$:

$$\mathcal{L} = -\sum_{i=0}^{N-1} y_i \log(x_i), \tag{6}$$

where N denotes the number of classes, and x_i and y_i represent the one-hot prediction and label at the i-th class, respectively. Considering the computation time and memory, we set the batch size to 32 and use the Adam optimizer to optimize the network parameters. In the training process, we start the learning rate at 0.01 and adaptively reduce the learning rate by half when the validation metric stagnates within 2 epochs. As shown in Fig. 11 b, c, both the training loss and validation loss converge to 0.006 and 0.1, while the learning rate decreases from 0.01 to 0.00001 after 200 epochs.

To verify the performance of the trained network, we first apply it to the validation dataset which are not included in training dataset. As shown in the Fig. 11 d, the predicted results are consistent with the labels. Besides, the predicted accuracy for five

https://doi.org/10.5194/essd-2024-337
Preprint. Discussion started: 7 October 2024

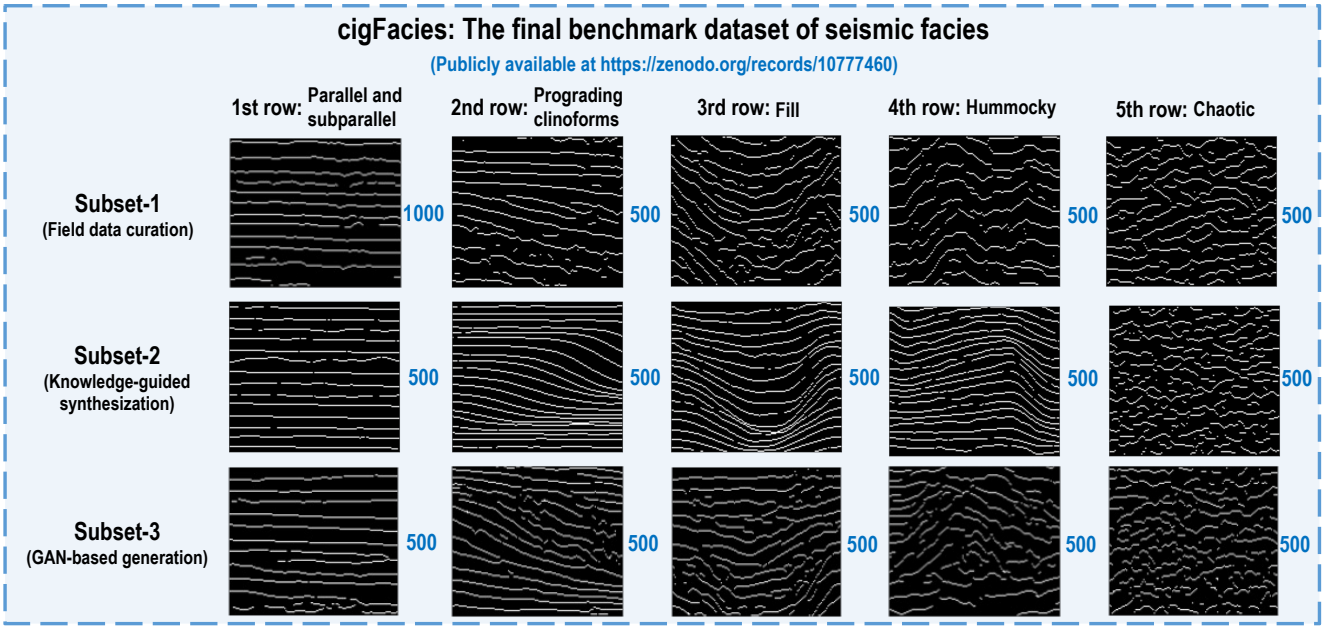

Figure 10. cigFacies: the final benchmark dataset of seismic facies construct from three strategies of field data curation, knowledge-guided synthesization and GAN-based generation. In this dataset, we totally generate 2000, 1500, 1500, 1500, and 1500 diverse facies samples for five common seismic facies (parallel and subparallel, prograding clinoforms, fill, hummocky and chaotic), respectively.

common seismic facies in validation dataset can up to 97.75%, 99%, 99.67%, 97.33% and 98.33%, which indicates that the trained network has successfully learned for automatic seismic facies classification.

3.2 Testing on the 3-D field seismic data

As shown in Fig. 12a, we develop a workflow for automatic seismic facies classification in 3-D field seismic data to further verify the performance of the trained network. We first use an automatic horizon-picking method (Wu and Fomel, 2018) to extract the top and bottom surfaces (green and red curves in Fig. 12b) of the target section in the 3-D field seismic data. Then we set a sliding window (blue box in Fig. 12b) bounded by the top and bottom surfaces to extract 2-D raw seismic image. Besides, each extracted 2-D seismic image is flattened with the bottom surface to eliminate the influence of the geological structures. We further apply the standardization and skeletonization processes to the flattened image to make it consistent with the training dataset. Finally, we feed the skeletonization image into the trained network for automatic seismic facies classification.

In this work, we apply the trained network on two distinct 3-D field seismic data (Longang and Yuanba) with complex geological structures. The Longgang (LG), Yuanba (YB) areas in Sichuan Basin develop a huge amount of platform margin reef complexes, which have emerged as an important field for oil and gas exploration (Chen et al., 2012; Xu et al., 2015; Tan et al., 2020). The first study case is the Permian Changhsing Formation of the LG 3-D seismic data shown in Fig. 12b and Fig. 13a. We employ the predicted workflow (Fig. 12a) with a sliding window traversing the entire 3D target strata,

https://doi.org/10.5194/essd-2024-337
Preprint. Discussion started: 7 October 2024

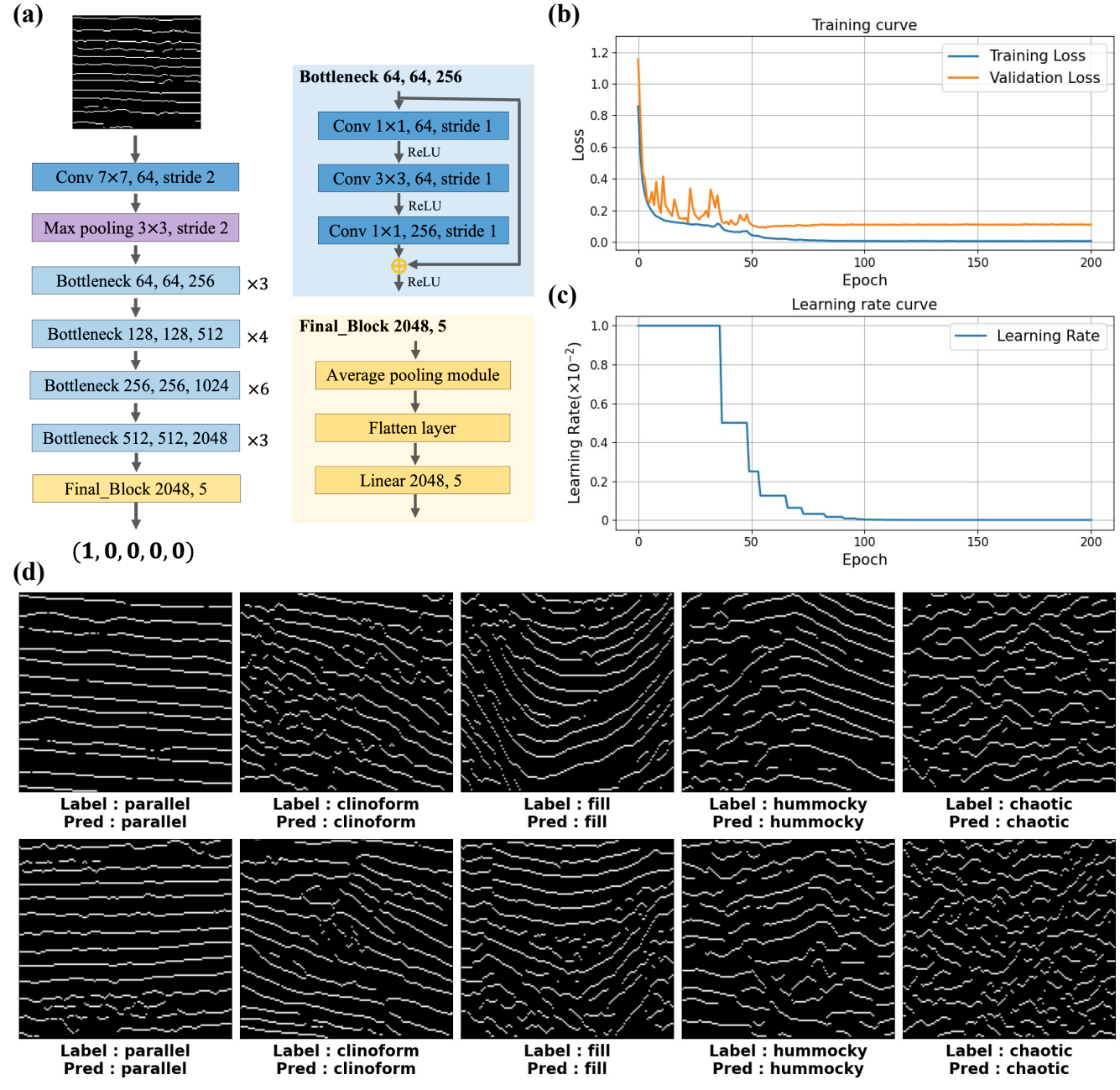

Figure 11. (a) The architecture of deep neural network (ResNet-50) used in this work for automatic seismic facies classification. The training (blue) and validation (orange) loss curves (b) and learning rate curve (c) during network training. After training the network, we apply the trained network to the validation dataset to verify its performance. The predicted results are consistent with the labels (d), which demonstrated that the trained network has successfully learned to automatically classify the seismic facies.

https://doi.org/10.5194/essd-2024-337
Preprint. Discussion started: 7 October 2024

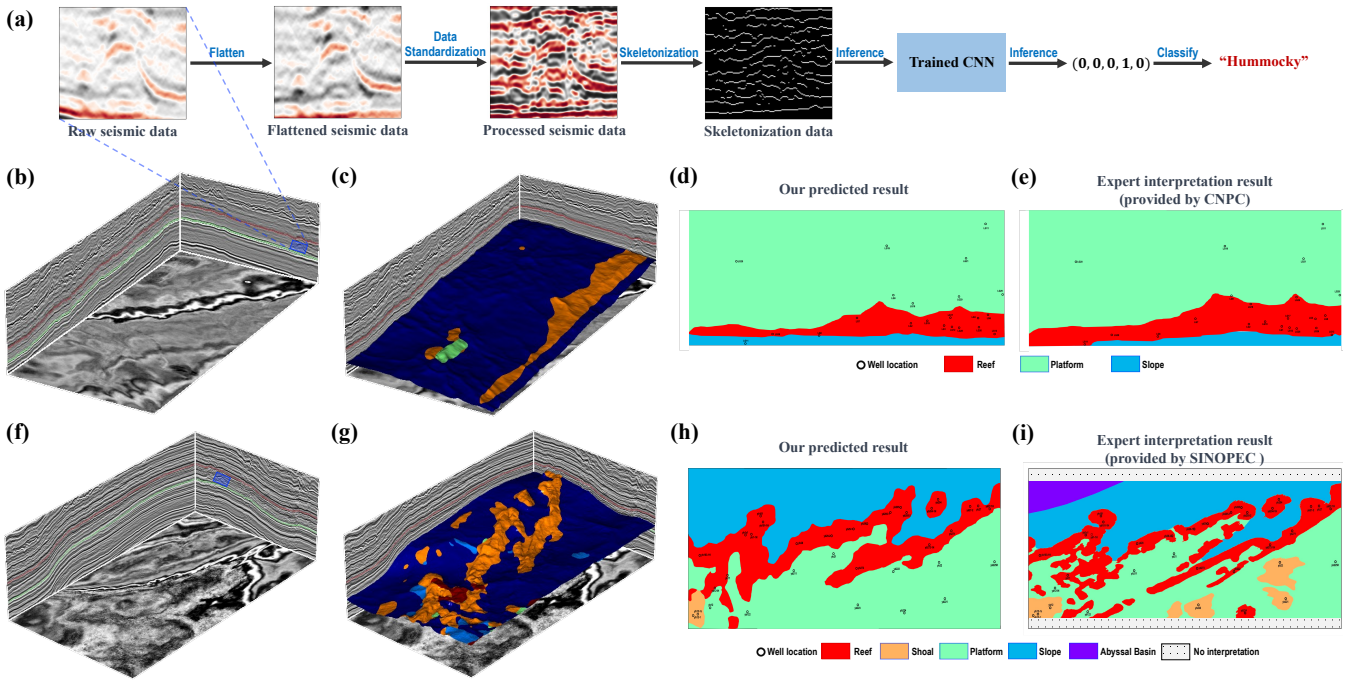

Figure 12. We employ the prediction workflow (a) with a sliding window scanning the entire 3D target section in the 3-D seismic data (b and f), yielding the seismic facies classification results (c and g). Then we obtain the corresponding sedimentary facies results (d and h) based on the predicted seismic facies result, well log informations, seismic data and geological and geophysics knowledge. Compared to the expert interpretation results (e and i), our predicted sedimentary facies results are high consistent.

yielding the seismic facies classification result shown in Fig. 12c. Besides, we display the predicted results with different 2-D profiles in Fig. 13b-f. The regions indicated by the blue arrows are correctly predicted to the hummocky facies, which are roughly consistent with geological structural uplift in corresponding 2-D seismic profiles. However, some artifacts or inaccurate predictions still appear in some areas indicated by red arrow in Fig. 13f, which is mainly due to the incomplete flattening of the strata. Finally, we obtain the corresponding sedimentary facies result (Fig. 12d) based on the predicted seismic facies result, well log informations, seismic data and geological and geophysics knowledge. Our final sedimentary facies result (Fig. 12d) is highly consistent with the expert interpretation of sedimentary facies shown in Fig. 12e.

The second study case is the Permian Changhsing Formation of the YB 3-D seismic data (Fig. 12f and Fig. 14a), which consists more complex geological structures than the LG 3-D seismic data. Following the same predicted workflow as before, we obtain the corresponding distributions of seismic facies and overlay the result with a manually interpreted horizon shown in Fig. 12g and Fig. 14b. The predicted distribution of hummocky seismic facies is consistent with the uplifted areas on the manually interpreted horizon. Such consistency can be better resolved in Fig. 14, where more 2D seismic profiles are displayed with the predicted result. However, some regions indicated by the red arrows (Fig. 14f) are incorrectly predicted as other seismic facies, which is probably due to the unsuitable scale of sliding window for the local regions, the influence of sliding

https://doi.org/10.5194/essd-2024-337
Preprint. Discussion started: 7 October 2024

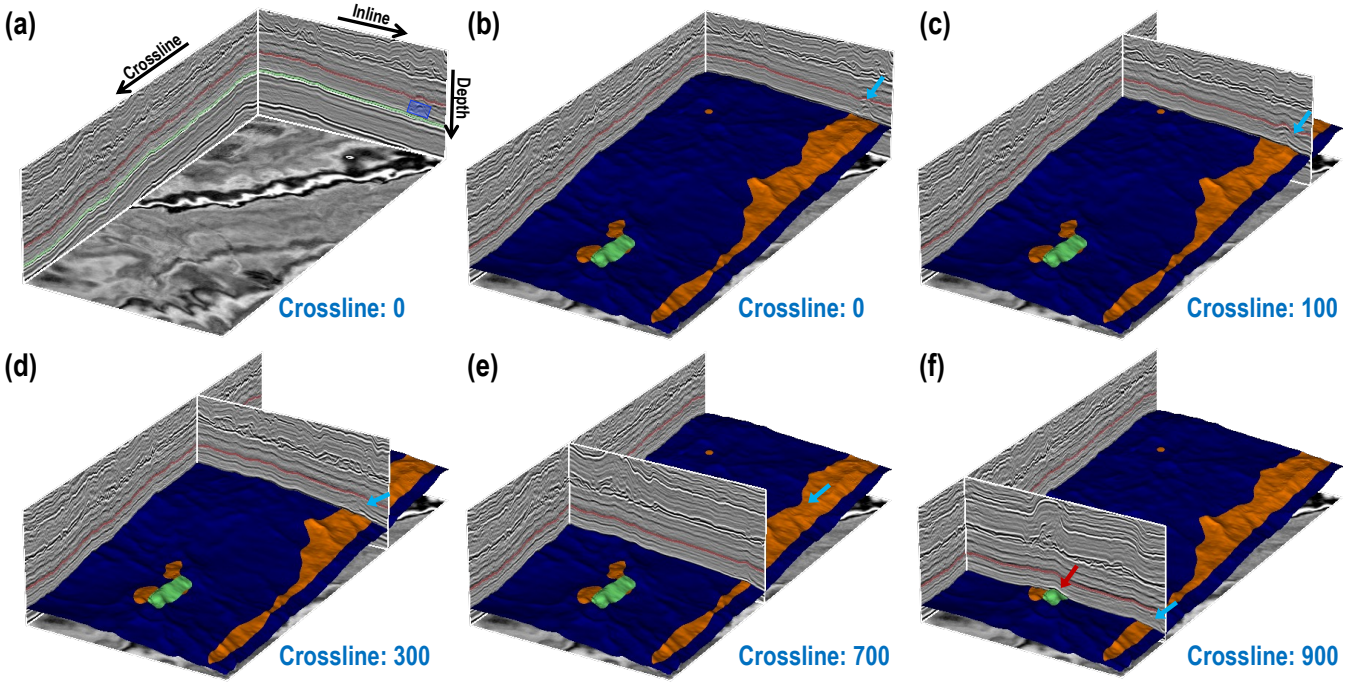

Figure 13. LG

window boundary effects, and incomplete flattening of the strata. Finally, we also generate the corresponding sedimentary facies results shown in Fig. 12h, where the platform margin reef are clearly and reasonably resolved and the spatial distribution of the platform margin is highly consistent with the expert interpretation in Fig. 12i.

4 Discussion

Applications on the validation dataset and two 3-D field seismic data indicate that the CNN model trained by our benchmark dataset has promising performance and strong generalization for automatic seismic facies classification. The benchmark dataset of seismic facies, guided by the knowledge graph and constructed from three strategies of field seismic data curation, knowledge-guided synthesization and GAN-based generation, can avoid the problems such as sample imbalanced, poor diversity and weak realism that usually occur in the traditional dataset construction methods. Besides, the data standardization and skeletonization processes successfully mitigate all potential data uncertainties (not related to seismic facies) across diverse data sources. This enables a deep learning model trained by the dataset to be effectively applied to field data from various surveys, thereby enhancing its generalizability.

Although the predicted results are roughly consistent with the human interpretation results, some limitations remain in the prediction with a sliding window. The normal or reverse faults in the seismic data probably introduce some unreasonable geological structures when flattening the seismic data, thus resulting in inaccurate predicted results. The proper setting of the

https://doi.org/10.5194/essd-2024-337
Preprint. Discussion started: 7 October 2024

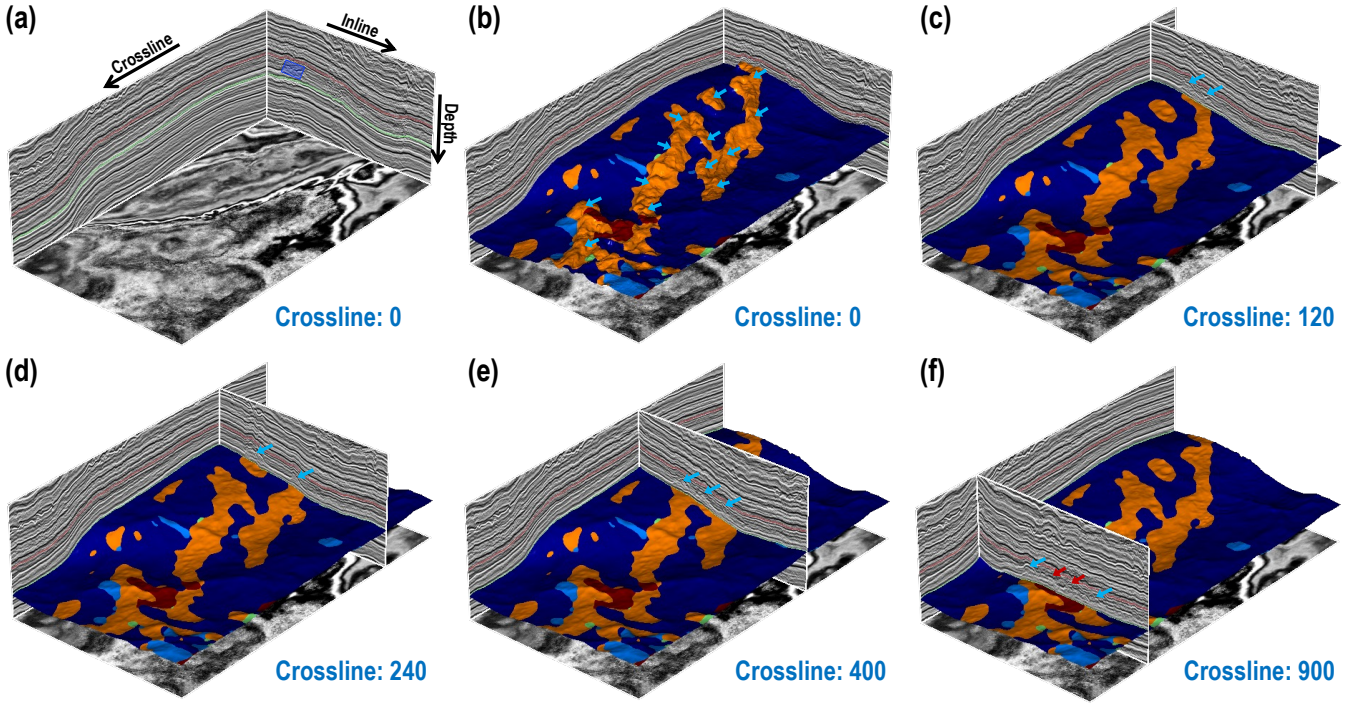

Figure 14. YB

sliding window size has a significant effect on the results, which needs to appropriately match the size of the key seismic facies in the seismic data. In addition, due to the predicted result is obtained by scanning pixel by pixel, some inaccurate predicted results may occur in the boundaries between different seismic facies, where the sliding window only contains the partial geological structure.

In the future, we can construct a more complete knowledge graph of seismic facies based on multiple parameters such as reflection configurations, continuity, amplitude, frequency, wave pattern and so on. We can further construct 3-D seismic datasets with multi-attribute features instead of 2-D skeletonization dataset that only contain geological structure informations. Additionally, we can develop a multi-scale 3D network for automatic seismic facies classification, which can improve the accuracy of predicted results in the boundaries between different seismic facies.

5 Conclusion

We have developed three strategies guided by a knowledge graph to build a benchmark dataset that is vast in scale, rich in features, and offers high realism. To the best of our knowledge, this dataset is the most extensive dataset of seismic facies currently available. The seismic facies knowledge graph, developed based on comprehensive literature review, summarizes various typical seismic facies types, along with their corresponding geological origins and seismic response features. This knowledge

https://doi.org/10.5194/essd-2024-337
Preprint. Discussion started: 7 October 2024

graph provides comprehensive guidance for the three strategies employed in building the benchmark dataset, ensuring the comprehensiveness and representativeness of the data sample construction. The first strategy of field seismic data curation yields the first subset that is authentic but exhibits some imbalance and limited diversity. The second strategy of sample synthesis, informed by the knowledge graph, generates a second subset of samples containing any category and pattern features, thereby addressing the issues of uneven sample type distribution and lack of diversity in the first subset. However, the synthesized samples also face the problem of being overly idealized and not sufficiently realistic. Consequently, a third strategy, based on AI generation, is adopted to refine the dataset construction. This strategy involves training a GAN model using the already constructed first and second subsets, then leveraging it to derive a third subset with diverse patterns and realistic features. By merging these three subsets, we have ultimately constructed a dataset containing 2000, 1500, 1500, 1500, and 1500 samples for five common seismic facies, respectively. This benchmark dataset has been demonstrated to effectively train a CNN model that achieves notable performance in seismic facies classification across two distinct 3-D field datasets. We have made this benchmark dataset publicly available, encouraging its further enhancement and utilization by others in the development and evaluation of deep learning approaches for seismic facies characterization.

6 Code and data availability

The benchmark dataset of seismic facies has been uploaded to Zenodo and are freely available at https://zenodo.org/records/10777460 (Gao et al., 2024a). The corresponding codes for constructing dataset and model training have been uploaded to Zenodo and are freely available at https://zenodo.org/records/13150879 (Gao et al., 2024b).

Author contributions. HG, XW, XS and MH initiated the idea of building the benchmark dataset of seismic facies and its application. HG, XW and XS initiated the idea of three strategies to constructing the benchmark dataset of seismic facies. HG, XW, XS, HS and HG conducted the first strategies of field data curation to build field facies samples. HG and XW tested and modified the code for the second and third strategies to build synthetic facies samples. HG carried out the experiments for the training and validation dataset. HG, HG and GW applying the trained network on the field seismic data. XW, XS and MH advised on the benchmark dataset preparation and predicted results analysis from a geological perspective. HG and XW prepared the paper, with contributions from all co-authors.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We thank the CNPC and SINOPEC for providing seismic data and expert interpretation results. We also thank the USTC supercomputing center for providing computational resources for this project.

https://doi.org/10.5194/essd-2024-337
Preprint. Discussion started: 7 October 2024

References

Chen, L., Lu, Y.-C., Guo, T.-L., and Deng, L.-S.: Growth characteristics of Changhsingian (Late Permian) carbonate platform margin reef complexes in Yuanba gas Field, northeastern Sichuan Basin, China, Geological Journal, 47, 524–536, 2012.

Duan, Y., Zheng, X., Hu, L., and Sun, L.: Seismic facies analysis based on deep convolutional embedded clustering, Geophysics, 84, IM87–IM97, 2019.

Dunham, M., Malcolm, A., and Welford, J.: Toward a semisupervised machine learning application to seismic facies classification, in: EAGE 2020 Annual Conference & Exhibition Online, vol. 2020, pp. 1–5, European Association of Geoscientists & Engineers, 2020.

Fensel, D., Simsek, U., Angele, K., Huaman, E., Kärle, E., Panasiuk, O., Toma, I., Umbrich, J., and Wahler, A.: Knowledge graphs, Springer, 2020.

Gao, H., Wu, X., Sun, X., and Hou, M.: cigFacies datasets: the massive-scale benchmark dataset of seismic facies, https://doi.org/10.5281/zenodo.10777460, 2024a.

Gao, H., Wu, X., Sun, X., and Hou, M.: cigFacies codes: cigFaciesNet for data generation and model training, https://doi.org/10.5281/zenodo.13150879, 2024b.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C.: Improved training of wasserstein gans, Advances in neural information processing systems, 30, 2017.

He, K., Zhang, X., Ren, S., and Sun, J.: Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., et al.: Knowledge graphs, ACM Computing Surveys (Csur), 54, 1–37, 2021.

Hu, X., Xu, Y., Ma, X., Zhu, Y., Ma, C., Li, C., Lü, H., Wang, X., Zhou, C., and Wang, C.: Knowledge System, Ontology, and Knowledge Graph of the Deep-Time Digital Earth (DDE): Progress and Perspective, Journal of Earth Science, 34, 1323–1327, 2023.

Jia, C. and Zhao, W.: Geological theory and exploration technology for lithostratigraphic hydrocarbon reservoirs, Petroleum Exploration and Development, 34, 257, 2007.

Karras, T., Aila, T., Laine, S., and Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation, arXiv preprint arXiv:1710.10196, 2017.

Li, J., Wu, X., Ye, Y., Yang, C., Hu, Z., Sun, X., and Zhao, T.: Unsupervised contrastive learning for seismic facies characterization, Geophysics, 88, WA81–WA89, 2023.

Liu, J., Dai, X., Gan, L., Liu, L., and Lu, W.: Supervised seismic facies analysis based on image segmentation, Geophysics, 83, O25–O30, 2018.

Liu, M., Jervis, M., Li, W., and Nivlet, P.: Seismic facies classification using supervised convolutional neural networks and semisupervised generative adversarial networks, Geophysics, 85, O47–O58, 2020.

Ma, C., Kale, A. S., Zhang, J., and Ma, X.: A knowledge graph and service for regional geologic time standards, Geoscience Frontiers, 14, 101 453, 2023.

Mitchum Jr, R. M., Vail, P. R., and Sangree, J. B.: Seismic stratigraphy and global changes of sea level: Part 6. Stratigraphic interpretation of seismic reflection patterns in depositional sequences: Section 2. Application of seismic reflection configuration to stratigraphic interpretation, AAPG Bulletin, 1977a.

https://doi.org/10.5194/essd-2024-337
Preprint. Discussion started: 7 October 2024

Mitchum Jr, R. M., Vail, P. R., and Thompson III, S.: Seismic stratigraphy and global changes of sea level: Part 2. The depositional sequence as a basic unit for stratigraphic analysis: Section 2. Application of seismic reflection configuration to stratigraphic interpretation, AAPG Bulletin, 1977b.

Paulheim, H.: Knowledge graph refinement: A survey of approaches and evaluation methods, Semantic web, 8, 489–508, 2017.

Puzyrev, V. and Elders, C.: Unsupervised seismic facies classification using deep convolutional autoencoder, Geophysics, 87, IM125–IM132, 2022.

Qi, J., Lin, T., Zhao, T., Li, F., and Marfurt, K.: Semisupervised multiattribute seismic facies analysis, Interpretation, 4, SB91–SB106, 2016.

Qian, F., Yin, M., Liu, X.-Y., Wang, Y.-J., Lu, C., and Hu, G.-M.: Unsupervised seismic facies analysis via deep convolutional autoencoders, Geophysics, 83, A39–A43, 2018.

Sangree, J. and Widmier, J.: Seismic stratigraphy and global changes of sea level: Part 9. Seismic interpretation of clastic depositional facies: Section 2. Application of seismic reflection configuration to stratigraphic interpretation, AAPG Bulletin, 62, 752–771, https://doi.org/10.1306/C1EA4E46-16C9-11D7-8645000102C1865D, 1977.

Sheriff, R.: Inferring stratigraphy from seismic data, AAPG Bulletin, 60, 528–542, 1976.

Tan, L., Liu, H., Tang, Y., Luo, B., Zhang, Y., Yang, Y., Liao, Y., Du, W., and Yang, X.: Characteristics and mechanism of Upper Permian reef reservoirs in the eastern Longgang Area, northeastern Sichuan Basin, China, Petroleum, 6, 130–137, 2020.

Veeken, P. C.: Seismic stratigraphy, basin analysis and reservoir characterisation, Elsevier, 2006.

Wrona, T., Pan, I., Gawthorpe, R. L., and Fossen, H.: Seismic facies analysis using machine learning, Geophysics, 83, O83–O95, 2018.

Wu, X. and Fomel, S.: Least-squares horizons with local slopes and multigrid correlations, Geophysics, 83, IM29–IM40, 2018.

Xu, G. and Haq, B. U.: Seismic facies analysis: Past, present and future, Earth-Science Reviews, 224, 103 876, 2022.

Xu, G., Xie, G., Long, K., and Song, X.: Sedimentary features and exploration targets of Middle Permian reservoirs in the SW Sichuan Basin, Natural Gas Industry B, 2, 415–420, 2015.

Zhang, H., Chen, T., Liu, Y., Zhang, Y., and Liu, J.: Automatic seismic facies interpretation using supervised deep learning, Geophysics, 86, IM15–IM33, 2021.

Zhang, L., Hou, M., Chen, A., Zhong, H., Ogg, J. G., and Zheng, D.: Construction of a fluvial facies knowledge graph and its application in sedimentary facies identification, Geoscience Frontiers, 14, 101 521, 2023.

Zhao, T.: Seismic facies classification using different deep convolutional neural networks, in: SEG International Exposition and Annual Meeting, pp. SEG–2018, SEG, 2018.

Zhao, T., Li, F., and Marfurt, K. J.: Seismic attribute selection for unsupervised seismic facies analysis using user-guided data-adaptive weights, Geophysics, 83, O31–O44, 2018.

Zhou, C., Wang, H., Wang, C., Hou, Z., Zheng, Z., Shen, S., Cheng, Q., Feng, Z., Wang, X., Lv, H., et al.: Geoscience knowledge graph in the big data era, Science China Earth Sciences, 64, 1105–1114, 2021.