Neural network-based CO_2 interpretation from 4D Sleipner seismic images

Bei Li[1]and Yunyue Elita Li[1]

[1]Department of Civil and Environmental Engineering, National University of Singapore, 117576 Singapore

Key Points:

- We train a 3D U-Net for an automatic end-to-end mapping from 4D seismic images to 3D CO_2 distribution.
- Successful applications of the trained neural network demonstrate its robustness and consistency.
- We analyze the neural network interpretation standards and provide training strategy for 2D sparse labels.

Corresponding author: Yunyue Elita Li, elita.li@nus.edu.sg

Abstract

Time-lapse or 4D seismic survey is a crucial monitoring tool for CO_2 geological sequestration. Conventional time-lapse interpretation provides detailed characterization of CO_2 distribution in the storage unit. However, manual interpretation is labour-intensive and often inconsistent throughout the long monitoring history, due to the inevitable changes in seismic acquisition and processing technology and interpreter's subjectivity. We propose a neural network (NN)-based interpretation method that translates baseline and monitoring seismic images to the probability of CO_2 presence. We use a simplified 3D U-net, whose training, validation and testing are all based on the Sleipner CO_2 storage project. The limited labels for training are derived from the interpreted CO_2 plume outlines within the internal sandstone layers for 2010. Then we apply the trained NN on different time-lapse seismic datasets from 1999 to 2010. The results suggest that our NN-based CO_2 interpretation has the following advantages: (1) high interpretation efficiency by automatic end-to-end mapping; (2) robustness against the processing-induced mismatch between the baseline and time-lapse inputs, relaxing the baseline reprocessing demands when compared to newly acquired or reprocessed time-lapse datasets; and (3) inherent interpretation consistency throughout multiple vintage datasets. Testing results with synthesized time-lapse images unveil that the NN takes both amplitude difference and structural similarity into account for CO_2 interpretation. We also compare 2D and 3D U-Nets under the scenario of sparse 2D labels for training. The results suggest that the 3D U-Net provides more continuous interpretation at the cost of larger computational resources for training and application.

Plain Language Summary

To reduce the greenhouse effect, captured and compressed CO_2 can be injected into subsurface area with special geological settings that permanently accommodate the greenhouse gas. Many pilot projects for such geological CO_2 sequestration have been ran for decades, during which various monitoring techniques have been experimented to study the interactions between injected CO_2 and the storage unit mainly for the purpose of evaluating storage safety. Such studies suggest that time-lapse 3D seismic survey is the key tool for detailed understanding of CO_2 behavior along time. However, the recorded seismic data must go through rigorous processing route to match the baseline and time-lapse datasets, so that the differences can be detected and interpreted by human labours. In this study, we propose to use the neural network (NN) to facilitate the human interpretation. It offers an efficient end-to-end mapping directly from the 4D seismic images to 3D CO_2 distribution. By incorporating differently processed data during training, the NN gains robustness against moderate mismatch between the baseline and time-lapse images. The generalized applications of the trained NN on different time-lapse data show great consistency throughout the monitoring history, which provides reliable analysis for CO_2 plume development as a function of time.

1 Introduction

Excessive emission of greenhouse gases due to anthropogenic activities has significantly contributed to climate change since the industrial revolution (Bachu & Adams, 2003). With the increase of fossil fuel consumption, CO_2 is responsible for more than 64% of the enhanced "greenhouse effect" (Bryant et al., 1997). Hence, CO_2 capture and storage (CCS) becomes an important measure for greenhouse gas mitigation, among which Geological CO_2 Sequestration (GCS) aims at removing CO_2 from the atmosphere by keeping them underground with suitable geomechanical conditions (Castelletto et al., 2013). During GCS projects, long-term monitoring is necessary for the purposes of understanding CO_2 behaviour in the reservoir, detecting CO_2 leakage from the storage unit, and assessing effects of contingency measures in case of leakage (Furre et al., 2017). Serving

for various monitoring purposes, time-lapse or 4D seismic survey is the most informative tool for detailed and quantitative CO_2 characterization in the storage complex varying along time (Boait et al., 2012; Bourne et al., 2014).

For GCS monitoring, time-lapse seismic analysis estimates the subsurface parameter changes between two separate seismic experiments due to CO_2 injection. The density and bulk modulus differences between the injected supercritical CO_2 and the originally saturated brine leads to dramatic velocity decrease in the storage reservoir, and consequently creates strong reflections at interfaces between the CO_2 and brine saturated rocks (R. Chadwick et al., 2005). Hence, conventional quantitative interpretations for CO_2 plume thickness and saturation are mainly based on velocity pushdown and reflection amplitude (A. Chadwick et al., 2010). However, detailed analysis still requires tedious manual interpretation and the outcome heavily depends on reliable processing with relative amplitude preservation and satisfying match between the baseline and time-lapse seismic images. Furthermore, it is difficult to maintain the interpretation consistent throughout the long-term GCS project. To partially automate this process, stratigraphic inversion (Clochard et al., 2010) and full waveform inversion (FWI) (Romdhane & Querendez, 2014) have been utilized to build high-resolution models of elastic impedance or velocity, based on which the CO_2 plume can be interpreted more accurately and objectively. Nonetheless, the inversion-based interpretation requires satisfying initial models, which also rely on substantial manual interpretation, and the computational cost for 3D inversion is expensive.

Different from conventional interpretation, machine learning-based seismic interpretation, trained by labels from experienced interpreter or realistic model building, can deliver satisfying results with much higher efficiency, e.g., in seismic facies analysis (Wrona et al., 2018), or faults (Wu et al., 2019), salt bodies (Guillen et al., 2015) and horizon (Geng et al., 2020) detection. For CO_2 interpretation in CCS projects, machine learning should be more attractive due to its high efficiency and inherent consistency for repetitive measurements throughout the long-term monitoring history. Using synthetic data generated by flow simulation, rock physics modeling and acoustic wave equation modeling, Wang et al. (2020) applied different machine learning algorithms to predict CO_2 saturation from seismic attributes and other downhole measurements. Sinha et al. (2020) treated the CO_2 leakage detection as an anomaly detection problem from CO_2 injection rates and pressure data using various neural networks (NNs). These successful applications are limited to frequently repeated small-scale measurements, whereas large-scale investigations, e.g., time-lapsed 3D surface seismic survey, are rarely interpreted by machine learning.

In this study, we propose a NN-based CO_2 interpretation that offers an end-to-end mapping from 4D seismic images, consisting of the baseline and time-lapse pairs, to 3D probability of CO_2 distribution. We employ a simplified 3D U-net which has been successfully utilized for fault detection from 3D seismic images (Wu et al., 2019). We train it by the shared seismic dataset and the benchmark model from the Sleipner CCS project, which is the world first industrial offshore CCS project starting injection from 1996, and storing around 18.5 million tonnes CO_2 by 2020 (Williams & Chadwick, 2021). The trained NN is applied to different monitoring time-lapse datasets acquired and processed in different years compared with the originally processed baseline dataset. With a single TITAN RTX GPU (24 G), the runtime for training and validation is around 3 hours, while the application takes only few seconds to obtain a complete 3D volume of CO_2 distribution. The NN interpretation results show high resolution with valid consistency throughout all available datasets. In addition, the NN also exhibits reasonable robustness against processing-induced mismatch between the baseline and time-lapse images for the input, which can potentially alleviate the reprocessing demands for newly acquired or reprocessed time-lapse datasets. To understand the interpretation standards of the trained NN, we perform tests based on real and synthetic data samples. The outcome suggests

that our trained NN considers both amplitude difference and the seismic structural similarity for detecting CO_2 distribution. Incidentally, we discuss the situation when only sparse 2D labels exist. The comparison between 2D and 3D U-Nets reveals that the 3D U-Net is advantageous in terms of interpretation resolution and continuity, but requires much more computational resources.

The structure of this paper is as follows: first, we introduce the NN architecture; second, we address the issue of training dataset generation with attested interpretation labels, and present the training and validation processes; third, we test the trained NN on datasets acquired and processed in different years; finally, we discuss the NN interpretation standards and provide a guideline for the case of sparse 2D labels, before we conclude.

2 Simplified 3D U-Net for CO_2 interpretation

The CO_2 interpretation in this study aims at depicting 3D CO_2 distribution in the reservoir from 4D seismic images. Therefore, we consider it as an image segmentation task, which assigns ones to CO_2 saturated parts, while zeros to other parts, in the corresponding seismic image domain. Similar image segmentation has been achieved in fault identification from seismic images (Wu et al., 2019), using a simplified version of the original U-Net (Ronneberger et al., 2015). We make some adjustments on the fault-detection network for CO_2 interpretation as shown in Figure 1. Firstly, our network input has two channels accommodating both baseline and time-lapse seismic images for CO_2 interpretation. Such input structure is trying to mimic the conventional interpretation, where we identify CO_2 according to amplitude changes and velocity pushdown in the time-lapse image w.r.t the baseline image. The analysis (downward) and synthesis (upward) paths are almost identical to those used for fault segmentation. However, our network uses half the number of feature maps at the bottom of the U-Net, to reduce memory and computational cost while preserve the performance for CO_2 interpretation. In addition, we also add an extra layer of batch normalization (BN) (Ioffe & Szegedy, 2015) before each rectified linear unit (ReLu), for faster convergence, as suggested by (Çiçek et al., 2016). The kernel size for max pooling in the downward layer is $2 \times 2 \times 2$ with a stride of 2 along each dimension. Correspondingly, we use trilinear algorithm for upsampling with the scale factor of 2. The kernel size for the 3D convolutional layer is $3 \times 3 \times 3$, except for the output layer, where convolution kernel size becomes $1 \times 1 \times 1$. The final Sigmoid function outputs a 3D probability volume for CO_2 presence in the corresponding seismic image domain.

The dimension of 3D cubes for both input and output is $64 \times 64 \times 64$, with the voxel size of 25 m $\times$ 25 m $\times$ 8 ms along inline, crossline and traveltime directions, respectively. Hence, the receptive field of each voxel in the bottom feature map is 200m $\times$ 200 m $\times$ 64 ms, which provides a satisfying balance between the computational cost and feature detection for CO_2 interpretation. To avoid the influence of disparate amplitude ranges between multi-vintages with different processing routes, we normalize the input cubes of baseline and time-lapse images, respectively, i.e., each image cube is subtracted by its mean value and divided by its standard deviation before being concatenated and fed to the NN.

3 Training and validation

3.1 Datasets generation

For the training input, publicly shared Sleipner 4D seismic dataset (Equinor, 2020a) includes abundant pairs of baseline and time-lapse datasets that have gone through time-lapse processing in different years. Figure 2 shows the acquired and processed years of different datasets. The baseline data is acquired in 1994, before the CO_2 injection started

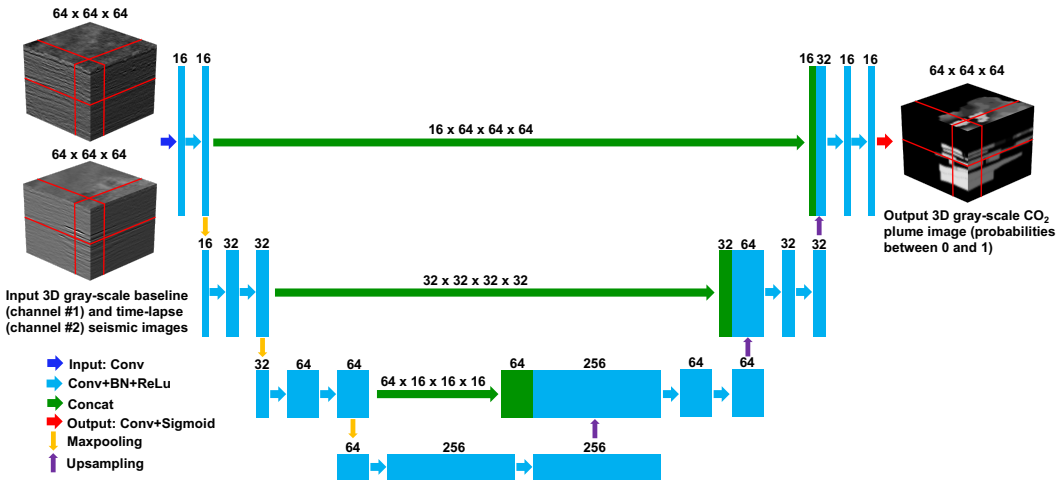

Figure 1. Simplified 3D U-Net for CO_2 interpretation from 4D seismic images.

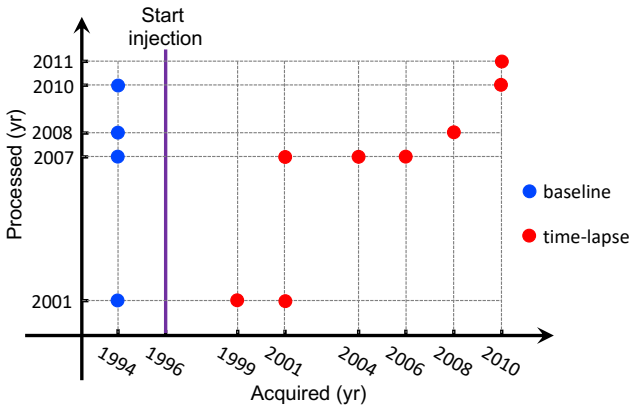

Figure 2. Acquiring and processing years for available seismic images in the 4D Sleipner seismic dataset.

in 1996 (Baklid et al., 1996). Then multiple time-lapse surveys have been repeated. In the shared datasets, the time-lapse data acquired in 1999, 2001, 2004, 2006, 2008 and 2010, are available. We refer to certain dataset as $xxpyy$, if it is acquired in the year of xx and processed in the year of yy. Each time-lapse dataset $xxpyy$ has a correspondingly reprocessed baseline dataset 94pyy, except for 10p11, which only has gone through image processing route without a matched baseline. Although all shared datasets contain near-, middle-, far-, and full-offset stacked images, we only utilize the near-offset part, which is adequate for the CO_2 interpretation in this study.

For the training label, it requires attested CO_2 interpretation corresponding to the input data. Here, we utilize the interpreted CO_2 plume boundaries of 2010 in nine internal sandstone layers provided in Sleipner 2019 Benchmark Model (Equinor, 2020b). Due to the lack of information for the CO_2 layers' thickness, we simply fill ones to the whole corresponding sandstone layer vertically within the plume lateral boundaries. Figure 3a illustrates the labeled CO_2 distribution in the model domain, along with the available interval velocity and reservoir interfaces from the benchmark model as well. Using these information, we convert the CO_2 label along depth into the image domain along traveltime (Figure 3b). The generated label can be seen as 3D probability volume of CO_2 distribution with lateral resolution of 100 m (as indicated in the benchmark model), and

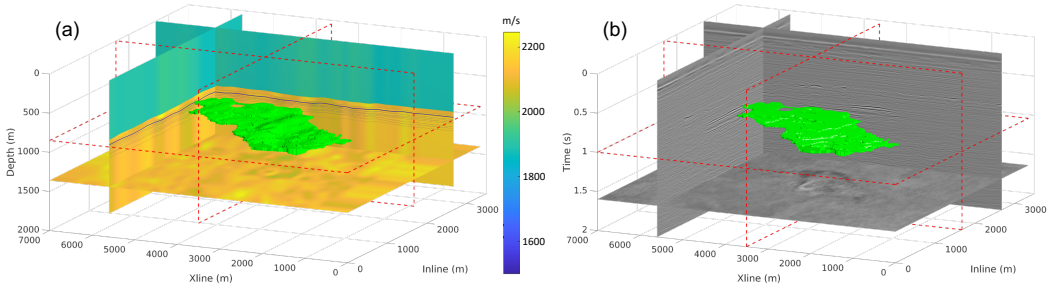

Figure 3. CO_2 plume labels for 2010 in (a) the model domain and (b) the image domain. The green blobs represent the CO_2 plume mask in 3D model or image domain. The red dashed squares indicate the sampled slice positions in 3D. The background color of the slices in (a) represent the interval velocity, while the wiggles represent the depth of interfaces in Utsira formation. The slices shown in (b) are stacked near-offset images from 10p10.

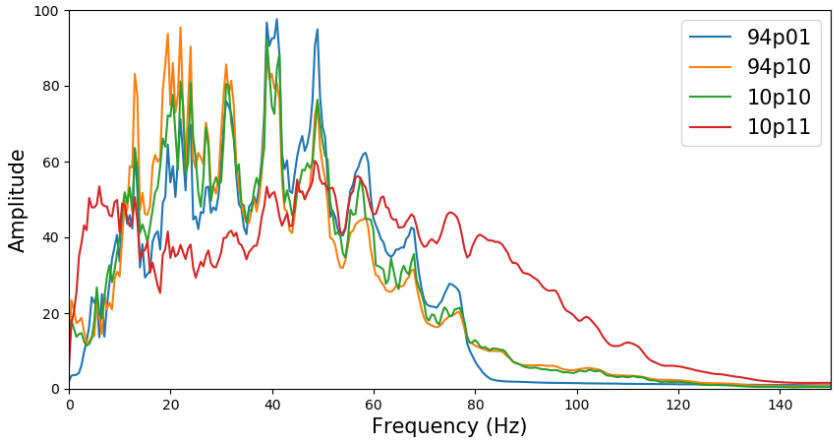

Figure 4. Comparison among amplitude spectra of 94p01, 94p10, 10p10, and 10p11.

vertical resolution of corresponding sandstone layers' thickness (in terms of corresponding two-way traveltime). Considering the interpretation uncertainties in the interval velocity and interface positions for the depth to traveltime conversion, we slightly smooth the generated image-domain label by a 3D Gaussian filter, so that the label margins are adjusted to lower confidence.

Since the labels are limited to 2010, it is natural to choose the baseline and time-lapse dataset pair for 2010, i.e., 94p10 and 10p10, to generate training inputs. However, we also include the originally processed baseline data 94p01 and the newly processed time-lapse data 10p11 in training dataset generation. Consequently, there are four types of combinations between baseline and time-lapse images as 94p01 vs. 10p10, 94p01 vs. 10p11, 94p10 vs. 10p10, and 94p10 vs. 10p11. Figure 4 shows the comparison among amplitude spectra of the four datasets. Reasonably, 94p10 and 10p10 are mostly identical to each other, and 94p01 is slightly different from them. However, 10p11 shows marked difference in the bandwidth compared to the others, since it has not gone through the time-lapse processing route. Hence, the various dataset combinations create sufficient processing discrepancy between the baseline and time-lapse inputs. Such discrepancy in the training dataset is necessary to improve the NN's tolerance against processing-caused mismatch between the baseline and the time-lapse inputs.

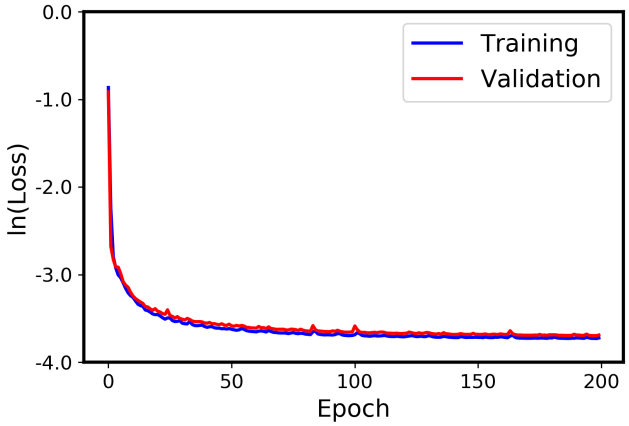

Figure 5. Training and validation losses in natural log scale.

We randomly sample 3D cubes correspondingly in one baseline, one time-lapse and the label volumes, which covers around 3.35×7.0 km on the surface and 2.0 s along traveltime. Originally, we have 720 cubes whose centers distributing randomly in the entire 3D volume. For each cube, the four combinations of baseline and time-lapse inputs corresponding to the same label are generated. Therefore, total 2880 samples are composed initially. Among these samples, around 60% show absolute zero CO_2 probability for their labels, which can significantly slow down the training convergence. Therefore, we reduce the number of zero CO_2 samples by randomly discarding some of them. Eventually, we have 1576 samples left, in which only 480 (around 1/3) of them are zero CO_2 samples. We divide these samples randomly into training and validation datasets with 1500 and 76 samples, respectively.

3.2 Training and validation

We use the binary cross-entropy (BCE) loss function since our labels are probability between zero and one. The total training epoch is 200. We use the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.0002. The batch size is 30 to diminish overfitting and improve the training efficiency with a single TITAN RTX GPU (24G). We use pytorch to implement the whole training and validation process and it takes approximately 3 hours.

Figure 5 shows the training and validation loss varying along the epoch number. Eventually, the decrease in the order of magnitude for both training and validation loss are around 2.8. In Figure 6, we show the NN predictions for two different samples from the training datasets. Comparing the baseline and time-lapse images in the samples, we observe good similarity for the area without CO_2 distribution. On the contrary, some amplitude anomalies indicate the CO_2 area in the corresponding labels. The NN predictions for both samples are quite consistent with the labels, despite of slightly lower resolution. More epochs or smaller batch size could further improve the resolution, but the robustness and generalization of the NN will probably be compromised.

4 Applications

4.1 Robustness test

After our NN is trained based on random samples from the seismic images and corresponding CO_2 labels for 2010, we firstly apply it to the same 4D seismic datasets, but

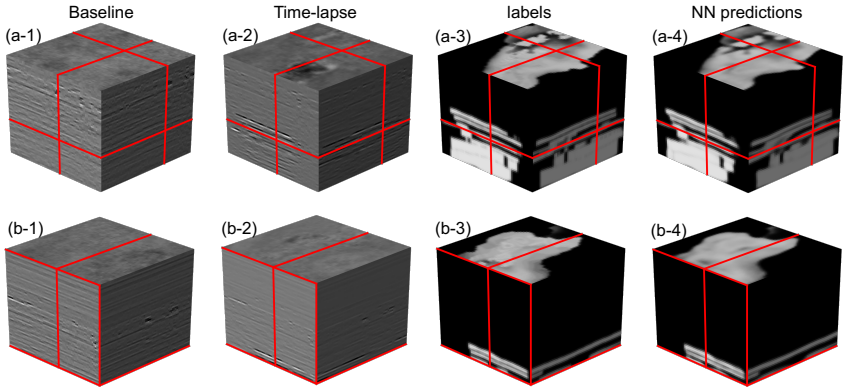

Figure 6. Trained NN predictions for two samples from the training dataset. The first two columns are the baseline and time-lapse seismic images as the NN input, and the last two columns are the human interpreted labels and NN predictions. The red lines indicate the slice positions shown by the cube surfaces.

with regularly sampled cubes on the entire 3D volume. The sampling number is 3×5×6 along inline, crossline and traveltime directions, resulting in 90 samples with approximately 40% overlapping in 3D. The runtime for such 90-sample test takes only few seconds. The obtained NN predictions for the 90 samples can be reconstructed into the original seismic image dimension through weighted summation. The 3D weighting function is:

$$w(x,y,t) = f(x)f(y)f(t), \tag{1}$$

where x, y and t represent inline, crossline and traveltime directions, respectively; f denotes a 1D weighting function with ones in the middle and gradually decaying to zero at the edges using Hanning window as follows:

$$f(a) = \begin{cases} 1, & |a| \le \frac{1}{2}\alpha L \\ \cos^2(\frac{\pi}{L}(|a| - \frac{1}{2}\alpha L)), & \frac{1}{2}\alpha L < |a| \le \frac{L}{2} \\ 0, & |a| > \frac{L}{2}, \end{cases} \tag{2}$$

$$\tag{3}$$

where $L = 64$ is the valid length of the weighting function which is consistent to the NN output size along each dimension, $\alpha = 0.6$ indicates the portion of ones w.r.t L in the middle of the weighting function. Thus, the weighted summation of the 90 NN predictions is

$$M(x,y,t) = \frac{\sum_{i=1}^{90} p_i w(x-x_i, y-y_i, t-t_i)}{\sum_{i=1}^{90} w(x-x_i, y-y_i, t-t_i)}, \tag{4}$$

where M is the reconstructed 3D CO_2 prediction, p_i is the NN prediction for the ith sample, whose cube center is (x_i, y_i, t_i).

To test the NN robustness against processing-caused mismatch between baseline and time-lapse inputs, we apply our NN on the four different sets of 90 samples corresponding to the four different combinations as 94p01 vs. 10p10, 94p01 vs. 10p11, 94p10 vs. 10p10, and 94p10 vs. 10p11, respectively. Figure 7 shows the reconstructed CO_2 distributions based on different NN predictions. We can see that the predicted CO_2 distributions for different combinations of baseline and time-lapse images are almost identical to the label (Figure 7e). Further calculating the BCE loss between each reconstructed prediction and the label, Table 1 shows that 94p10 vs. 10p10 provides the best result,

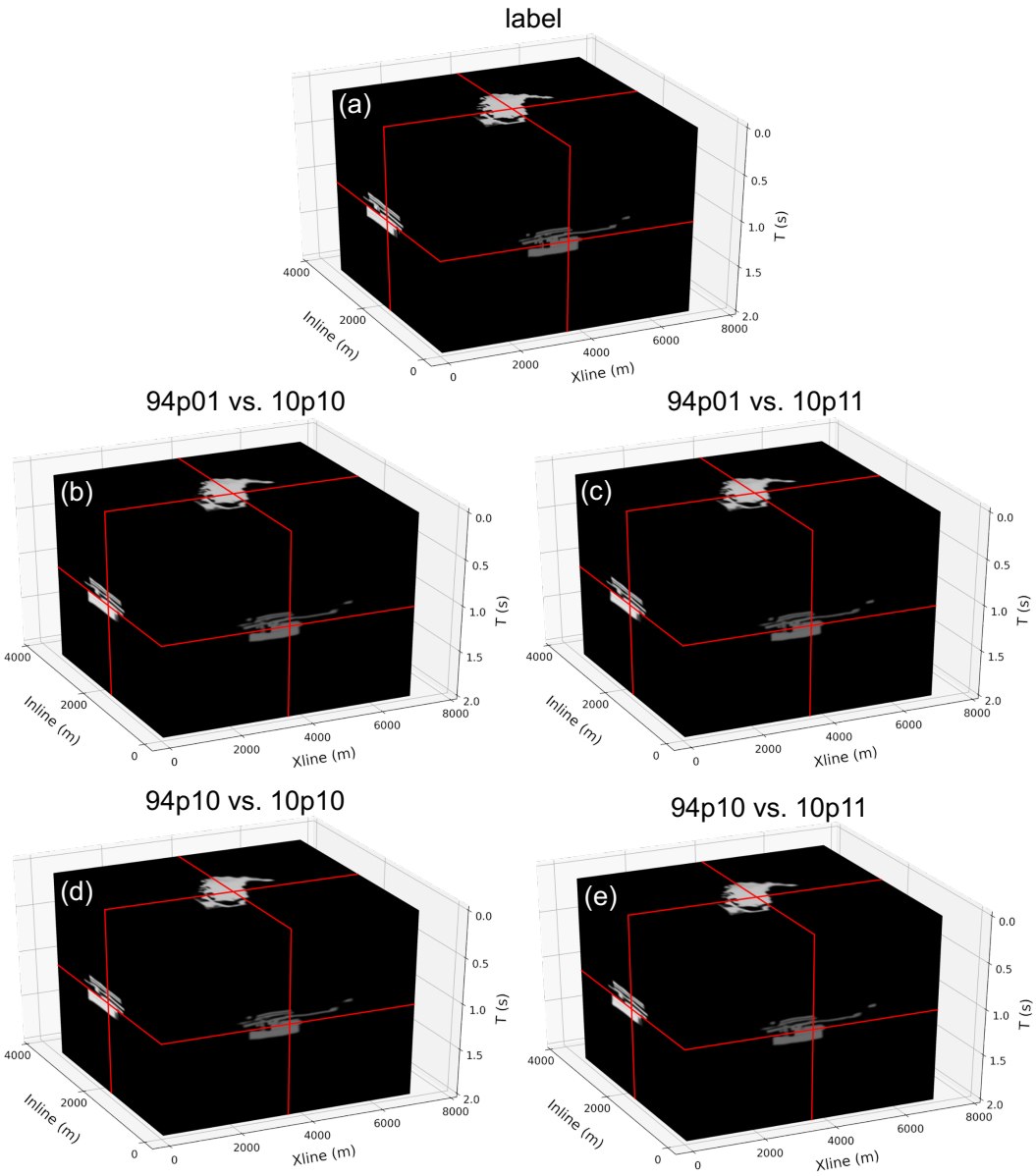

Figure 7. Comparison between (a) the label and (b,c,d,e) the reconstructed CO_2 distributions based on NN predictions using four different combinations of baseline and time-lapse images.

and using 10p11 as time-lapse input always leads to larger losses in comparison with 10p10. Such observations are consistent with the dataset similarity represented by the spectra comparison shown in Figure 4. Regardless of the insignificant differences in the BCE losses, the NN predictions are visually undifferentiated, reflecting strong robustness of the trained NN against moderate processing mismatch between the baseline and time-lapse images.

In Figure 8, we further compare the absolute amplitude anomaly with the human interpreted and NN predicted CO_2 distributions in the top sand wedge layer above the Utsira Formation. The 2D amplitude anomaly is obtained by calculating the vertical mean of the absolute amplitude difference between 94p10 and 10p10 within the designated layer, whereas the interpreted CO_2 distributions are the vertical means of the 3D CO_2 probabilities within the same layer. We can see that the amplitude anomaly does provide a

Table 1. BCE loss of the reconstructed CO_2 distributions w.r.t the label.

BCE loss: time-lapse \ baseline	94p01	94p10
10p10	$4.86e^{-3}$	$4.85e^{-3}$
10p11	$4.91e^{-3}$	$4.93e^{-3}$

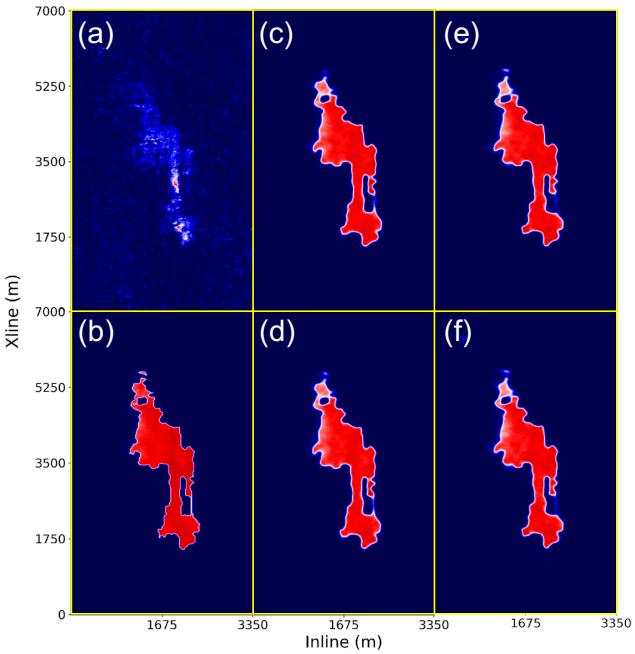

Figure 8. Comparison between (a) the absolute amplitude anomaly with (b) the human interpreted and (c,d,e,f) NN predicted CO_2 distributions in the top sand wedge layer above the Utsira Formation. For baseline input, (c) and (e) use 94p01, while (d) and (f) use 94p10; for time-lapse input, (c) and (d) use 10p10, while (e) and (f) use 10p11.

rough clue for the CO_2 distribution. However, it requires further processing and more detailed analysis to obtain the high-resolution CO_2 plume as shown in the label (Figure 8b). Contrarily, the trained NN achieves accurate CO_2 depiction with high resolution directly from the baseline and time-lapse images, even when noticeable processing mismatch exists.

4.2 Consistency test

To further generalize our trained NN, we apply it to other available 4D seismic vintages shared by the Sleipner CO_2 storage project. Since we have proved the robustness of our trained NN, we use the same originally processed 94p01 as the NN baseline input for all time-lapse inputs: 99p01, 01p01, 04p07, 06p07, 08p08, and 10p10. Figure 9 displays the NN interpreted CO_2 distribution in the top (L9), middle (L5) and base (L1) of the internal sandstone layers in Utsira Formation, developing from 1999 to 2010. In all displayed layers, the NN predictions are reasonably compacted with clear and continuous boundaries. Moreover, they are growing steadily throughout the decade, although the CO_2 plume in L1 expands noticeably slower, due to the buoyancy of supercritical

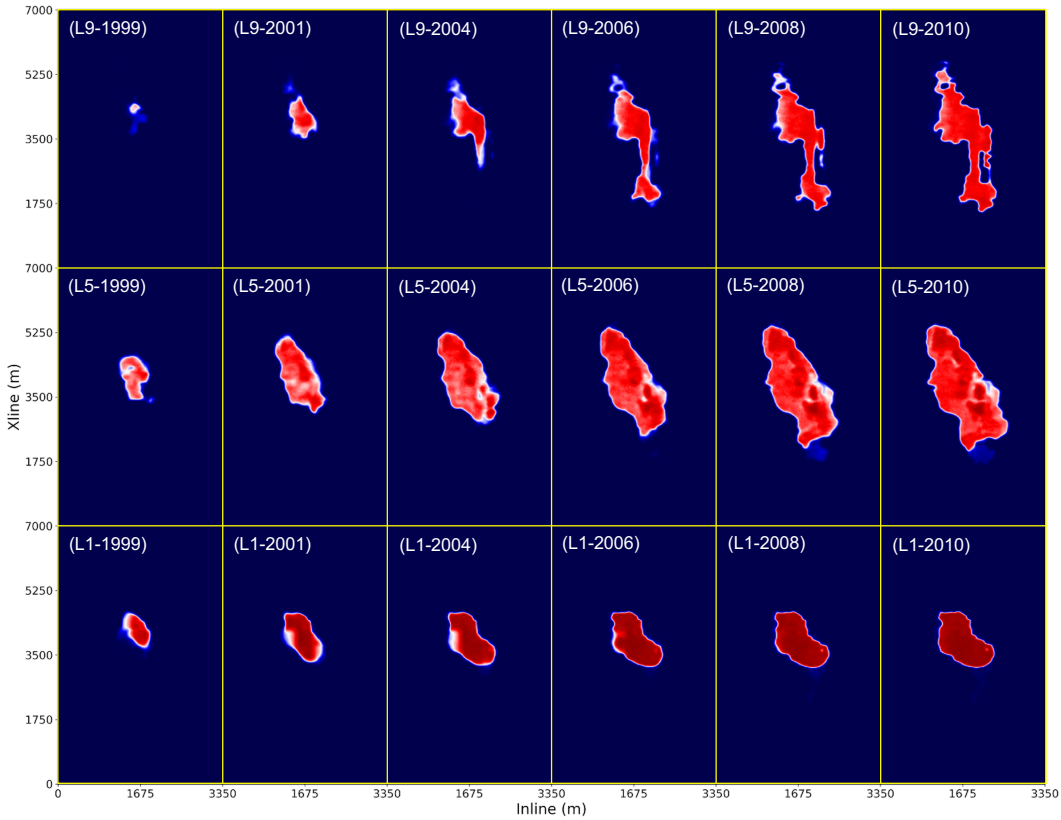

Figure 9. NN interpreted CO_2 plume expanding along L9 (top), L5 (middle) and L1 (base) of the Utsira Formation, from 1999 to 2010.

CO_2 in the saline aquifer (Arts et al., 2008). In 1999, the top layer result (Figure 9(L9-1999)) shows a singularity on the probability map, indicating the injected CO_2 has just reached the top of the formation (A. Chadwick et al., 2010). Similar singularities are also visible in Figures 9(L9-2004), (L9-2008) and (L5-1999), suggesting that our NN interpretation has the potential for high-resolution leakage detection or feeder recovery. Finally, we can also identify the migration directions of CO_2 plume in different layers, e.g., in L5, the CO_2 plume is mainly lengthened along SW-NE directions, and specifically, towards the NE direction since 2004. Generally, the NN interpretations along time are reasonably consistent in terms of CO_2 migration and plume expansion in the storage unit.

5 Discussion

5.1 Analysis for NN interpretation standards

We design a test to qualitatively explain how does the trained NN determine the CO_2 distribution given baseline and time-lapse inputs. Direct observation from Figure 6 indicates that large amplitude anomaly in the time-lapse image w.r.t the baseline image is the apparent key. In view of this intuitive hypothesis, we sample two cubes whose centers are both in the inline assemble at 1625 m from 94p10 and 10p10 as shown in Figure 10. The first sample includes the primary reflections caused by CO_2 accumulation in the storage unit, and the second sample contains the corresponding surface-related multiples right below the first sampling position. Hence, there are amplitude anomalies in both samples, despite that their labels are totally different as shown in Figure 10c. We feed these two samples of baseline and time-lapse cubes (Figures 11a and b) into the

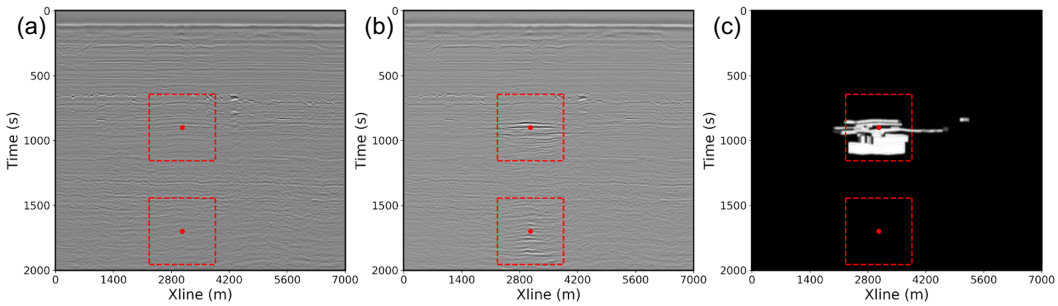

Figure 10. The inline assembles at 1625 m for (a) baseline and (b) time-lapse images from 94p10 and 10p10, along with the (c) CO_2 distribution label. The red dots and dashed lines indicate the centers and boundaries of the sampled cubes within the inline assemble.

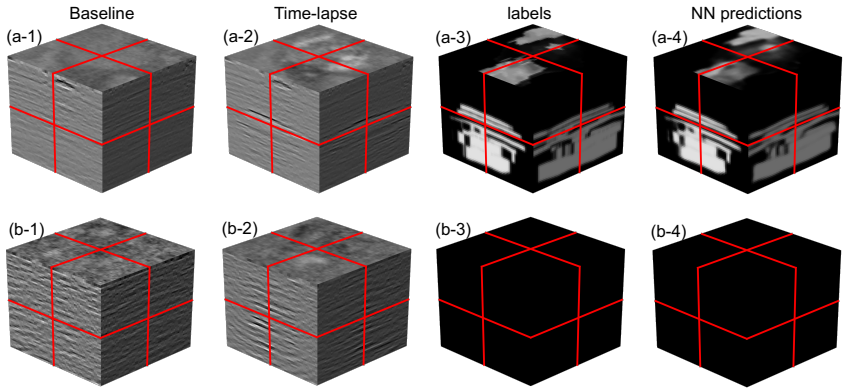

Figure 11. Trained NN predictions for the two samples shown in Figure 10. The first two columns are the baseline and time-lapse seismic images as the NN input, and the last two columns are the human interpreted labels and NN predictions.

trained NN, and the predictions are shown in Figure 11c, which are consistent with the labels (Figure 11d). This implies that the NN does not solely rely on the amplitude difference anomaly between the baseline and time-lapse images to determine the CO_2 distribution.

To further explore why the NN does not misinterpret the surface-related multiples, we modify the time-lapse image for the this sample. The modifications and corresponding NN predictions are shown in Figure 12. In the first modified sample, we scale up the center area of the original time-lapse image (Figure 11b-2) as the new time-lapse image (Figure 12c), whereas in the second modified sample, we scale up the center area of the original baseline image (Figure 11b-1) as the new time-lapse image (Figure 12e). Both modified samples now share the same baseline image as shown in Figure 12a and the locally scaling multiplier is displayed in Figure 12b, whose center area has the value of 4, whereas the outer area is 1. By feeding the modified samples to the trained NN, we obtain the predictions shown in Figures 12d and f. It appears that further increasing the amplitude of the multiples does create slight CO_2 probability in the prediction. However, when we directly scale up the baseline amplitude in specific area as the time-lapse image, significant CO_2 probability emerges in the scaled up area. Hence, it reveals that the trained NN also considers the structural similarity between the baseline and time-lapse images, in addition to the amplitude difference anomaly. This indicates that our trained NN performs similarly with human interpreter.

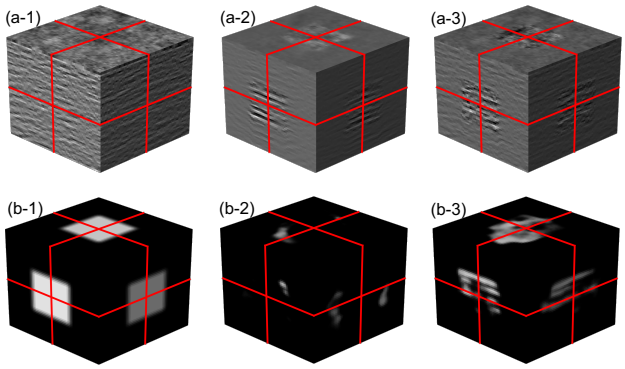

Figure 12. Modifications of the time-lapse images in the second sample shown in Figure 11. (a) is the common baseline; (b) is the locally scaling multiplier; (c) and (e) are new time-lapse images; (d) and (f) are NN predictions corresponding to (c) and (e), respectively.

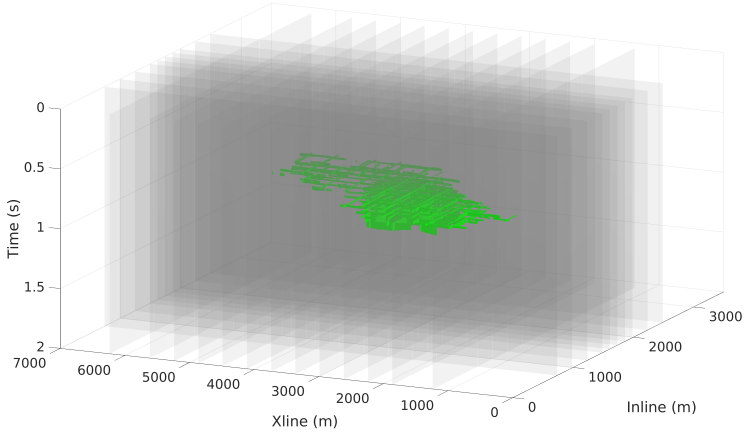

Figure 13. 2D slice labels for CO_2 distribution, indicated by the green patches on each slice represented by translucent surfaces.

5.2 2D vs. 3D NN using sparse 2D labels

The presented NN based on the Sleipner 4D seismic dataset is trained by a complete 3D label generated from CO_2 plume boundary interpretation for 2010. However, in more general cases, the interpreted labels for CO_2 distribution are often available in sparse 2D slices of inline and/or crossline assembles, due to human perception limitations. We create an example of interpreted slices along both inline and crossline directions shown in Figure 13. We sample 10 slices along inline direction, and 13 slices along crossline direction. The slice interval is smaller (125 m along inline direction and 375 m along crossline direction) near the storage unit, and larger (250 m along inline direction and 625 m along crossline direction) away from the storage unit, for a better representation of the label target.

One way of utilizing such sparse 2D labels is to directly train a 2D U-Net. Another way is to train the 3D U-Net with corresponding weight on the sparse labels during loss evaluation (Çiçek et al., 2016). Here, we compare these two strategies to offer a guideline under such realistic scenario of sparse interpretation labels.

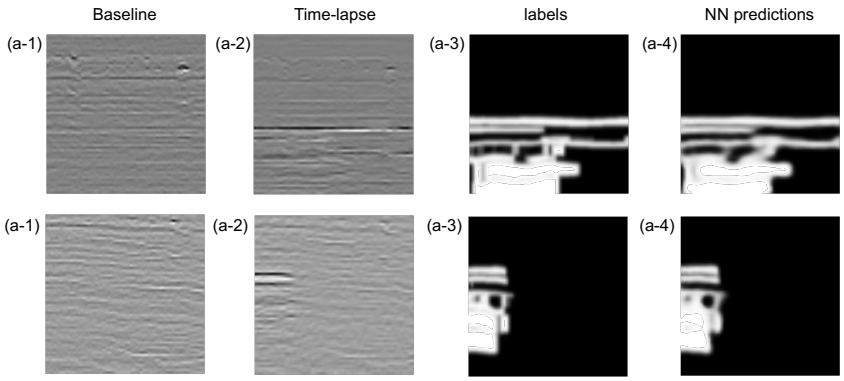

Figure 14. Trained 2D NN predictions for two samples from the training dataset with sparse 2D slice labels. The first two columns are the baseline and time-lapse seismic images as the NN input, and the last two columns are the human interpreted labels and NN predictions.

5.2.1 2D U-Net with sparse 2D labels

We use the same architecture shown in Figure 1 for the 2D U-Net, except we reduce the dimensionality from 3D to 2D for the convolution, max pooling and upsampling. To generate training dataset, we also utilize both 94p01 and 94p10 as the baseline input, along with 10p10 and 10p11 as the time-lapse input. Since trace intervals along both inline and crossline directions are the same, we can train a 2D U-Net applicable on both inline and crossline assembles. The 2D patch size is 64×64, with the grid size of 25 m× 8 ms along inline/crossline and traveltime directions, respectively. We sample 2D patches on all available slices shown in Figure 13 correspondingly in one baseline image, one time-lapse image and the label slice. The number of initially sampled patch centers is 1320, resulting in 5280 2D training samples. We also reduce the number of zero CO_2 samples, and eventually keep 1500 samples for training, of which around 25% are zero CO_2 samples. The 2-D U-Net training uses the same training parameters as for the 3D U-Net. Notice that due to the dimensionality reduction, the training time has reduced significantly from 3 hours to 5 minutes for 200 epochs using the same TITAN RTX GPU. Figure 14 shows the results of two training samples. It appears that the NN predictions are consistent with the corresponding 2D labels.

To test the trained 2D U-Net, we apply it on 94p01 vs. 04p07 and 94p01 vs. 10p10, respectively, by regularly sampling 2D patches along all inline and crossline assembles for the 3D dataset volume. The sampling numbers are also $(3, 5, 6)$ along inline, crossline and traveltime directions. Traversing all the inline and crossline assembles respectively, we have the total sampling number as $269 \times (5 \times 6) + 561 \times (3 \times 6) = 18168$, in which 269 and 561 are the number of inline and crossline assembles, respectively. We combine these 2D NN predictions together by the weighted summation as shown in Equation 4, only now with 2D Hanning-windowed weighting functions. Figures 16a and b display the reconstructed 3D CO_2 distributions for the two tests. Compared to the label and the predictions from the original 3-D U-Net shown in Figure 7, the 2010 result predicted by 2D U-Net shows lower resolution and more artifacts around the edges of the volume. Similar defects are also visible in the 2004 result obtained from 2D U-Net. We further display the reconstructed 2D CO_2 distributions along the top sand wedge layer for both tests in Figures 17a and b. Distinct scratch-like artifacts are visible in comparison with the corresponding 3D U-Net predictions shown in Figure 9. This is because the 2D U-Net cannot preserve the continuity along the 3rd dimension vertical to the plane where the 2D U-Net is applied. Although we apply the 2D U-Net along both inline and crossline assembles then combine the predictions by weighted summation, the outcomes simply

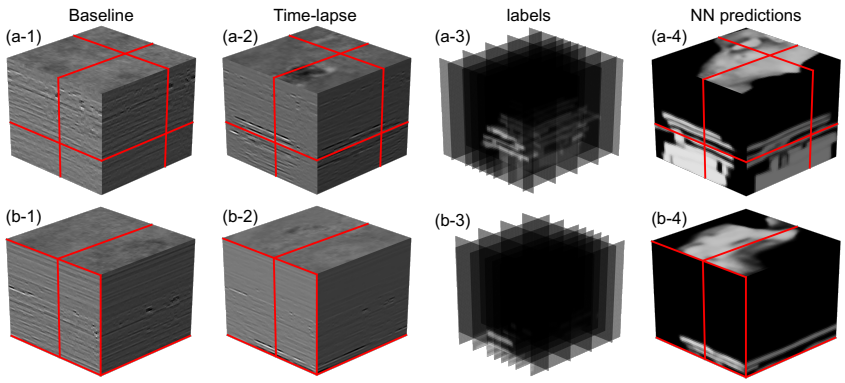

Figure 15. Trained 3D NN predictions for two samples from the training dataset with sparse 2D slice labels. The first two columns are the baseline and time-lapse seismic images as the NN input, and the last two columns are the human interpreted labels and NN predictions.

present discontinuities along both directions (the horizontal and vertical "scratches") as shown in Figures 17a and b.

5.2.2 3D U-Net with sparse 2D labels

We train a new 3D U-Net with sparse 2D labels using the same training dataset and parameters as those used in the original 3D U-Net with full 3D labels. Thus, for each sampled cube, the baseline and time-lapses images are unchanged, but the label now has a corresponding weight, in which the sampled 2D slice position is assigned as one, whereas other part is zero. The weight is implemented during the BCE loss calculation and then backpropagated to influence the 3D U-Net update. The runtime for 200 epochs are basically the same as for the original U-Net training. Figure 15 shows the same samples displayed in Figure 6. We can see the NN inputs are exactly the same, whereas the labels are vastly different, since now we only have certain vertical slices (Figures 15a-3 and b-3) instead of the whole cube (Figures 6a-3 and b-3). However, the NN predictions are reasonably consistent. The horizontal slices shown in Figures 15a-4 and b-4 have been retrieved with satisfying resolution and continuity, even though they are not exactly the same as predictions from NN trained by full 3D labels shown in Figures 6a-4 and b-4.

We also test the sparsely trained 3D U-Net on 94p01 vs. 04p07 and 94p01 vs. 10p10, respectively. The 3D reconstructed CO_2 distributions are shown in Figures 16c and d. It appears that the sparsely trained 3D U-Net results provide much higher resolution than those obtained from 2D U-Net (Figures 16a and b). Figures 17c and d further display the top sand wedge layer CO_2 distributions from the sparsely trained 3D U-Net. Compared to the 2D U-Net results (Figures 17a and b), there are no scratch-like artifacts and the plume boundaries exhibit more continuity.

In summary, the 3D U-Net trained by weighted sparse labels generally present higher-quality interpretation than the 2D U-Net trained by the same labels, in terms of resolution, boundary continuity and artifacts. However, the 3D U-Net training and applications require much more computational resources than the 2D U-Net. Hence, we suggest to use 3D U-Net even for sparse 2D labels as long as necessary computational power (GPU with large enough memory) is accessible.

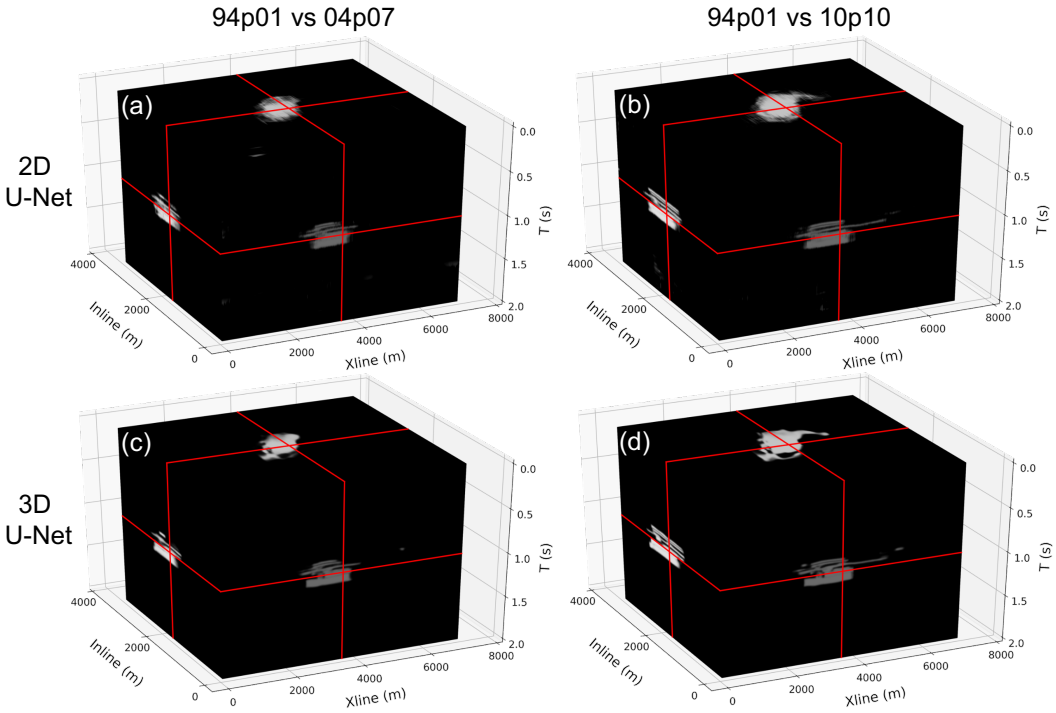

Figure 16. Reconstructed 3D CO_2 distributions from NN predictions for the tests of 94p01 vs. 04p07 and 94p01 vs. 10p10, using trained 2-D U-Net and 3-D U-Net, respectively.

6 Conclusion

We utilized a simplified 3D U-Net to interpret the 3D CO_2 distribution from large 4D seismic images. The NN is trained on Sleipner datasets acquired from 1994 and 2010, but processed in 2001, 2010 and 2011. Hence, the trained NN shows reasonable robustness against processing-caused mismatch between the baseline and time-lapse images. Moreover, the generalized applications on other time-lapse images acquired from 1999 to 2008 also achieve satisfying results with great interpretation consistency. We also analyzed the NN interpretation standards and provide NN training strategy under more realistic scenario where only sparse 2D labels are available. Overall, the studied 3D U-Net is proved to be an efficient, robust and flexible tool for CO_2 interpretation from 4D seismic monitoring datasets during long-term CCS projects.

Acknowledgments

The authors acknowledge the Singapore Economic Development Board for its financial support through the Petroleum Engineering Professorship. The reproducible codes and all supplementary materials related to this article can be found online `https://github.com/nusbei/CO2_Sleipner`.

References

Arts, R., Chadwick, A., Eiken, O., Thibeau, S., & Nooner, S. (2008). Ten years' experience of monitoring CO_2 injection in the Utsira Sand at Sleipner, offshore Norway. *First Break*, *26*(1).

Bachu, S., & Adams, J. (2003). Sequestration of CO_2 in geological media in response to climate change: capacity of deep saline aquifers to sequester CO_2 in

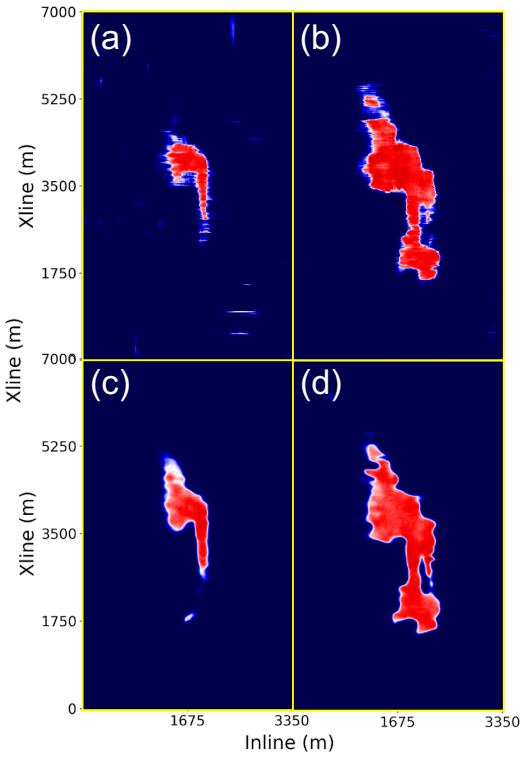

Figure 17. NN interpreted CO_2 distributions in the top sand wedge layer above the Utsira Formation. (a) and (b) are the 2D U-Net results from 94p01 vs. 04p07 and 94p01 vs. 10p10, respectively, while (c) and (d) are the 3D U-Net results from 94p01 vs. 04p07 and 94p01 vs. 10p10, respectively.

solution. *Energy Conversion and management*, *44*(20), 3151–3175.

Baklid, A., Korbol, R., Owren, G., et al. (1996). Sleipner Vest CO_2 disposal, CO_2 injection into a shallow underground aquifer. In *Spe annual technical conference and exhibition.*

Boait, F., White, N., Bickle, M., Chadwick, R., Neufeld, J., & Huppert, H. (2012). Spatial and temporal evolution of injected CO_2 at the Sleipner Field, North Sea. *Journal of Geophysical Research: Solid Earth*, *117*(B3), B03309.

Bourne, S., Crouch, S., & Smith, M. (2014). A risk-based framework for measurement, monitoring and verification of the Quest CCS project, Alberta, Canada. *International Journal of Greenhouse Gas Control*, *26*, 109–126.

Bryant, E., Bryant, E. A., & Edward, B. (1997). *Climate process and change.* Cambridge University Press.

Castelletto, N., Gambolati, G., & Teatini, P. (2013). Geological CO_2 sequestration in multi-compartment reservoirs: Geomechanical challenges. *Journal of Geophysical Research: Solid Earth*, *118*(5), 2417–2428.

Chadwick, A., Williams, G., Delepine, N., Clochard, V., Labat, K., Sturton, S., ... others (2010). Quantitative analysis of time-lapse seismic monitoring data at the Sleipner CO_2 storage operation. *The Leading Edge*, *29*(2), 170–177.

Chadwick, R., Arts, R., & Eiken, O. (2005). 4D seismic quantification of a growing CO_2 plume at Sleipner, North Sea. In *Geological society, london, petroleum geology conference series* (Vol. 6, pp. 1385–1399).

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In *International conference on medical image computing and computer-assisted intervention* (pp. 424–432).

Clochard, V., Delépine, N., Labat, K., & Ricarte, P. (2010). CO_2 plume imaging using 3D pre-stack stratigraphic inversion: A case study on the Sleipner field. *First Break*, *28*(1).

Equinor. (2020a). *Sleipner 2019 Benchmark Model.* Retrieved from `https://co2datashare.org/dataset/e6f67cbd-abf3-4d85-a118-ed386a994c2c` doi: 10.11582/2020.00004

Equinor. (2020b). *Sleipner 4D Seismic Dataset.* Retrieved from `https://co2datashare.org/dataset/cbdc354c-fa61-4ab4-a0b4-134e1350a82b` doi: 10.11582/2020.00005

Furre, A.-K., Eiken, O., Alnes, H., Vevatne, J. N., & Kiær, A. F. (2017). 20 years of Monitoring CO_2-injection at Sleipner. *Energy Procedia*, *114*, 3916–3926.

Geng, Z., Wu, X., Shi, Y., & Fomel, S. (2020). Deep learning for relative geologic time and seismic horizons. *Geophysics*, *85*(4), WA87–WA100.

Guillen, P., Larrazabal*, G., González, G., Boumber, D., & Vilalta, R. (2015). Supervised learning to detect salt body. In *Seg technical program expanded abstracts 2015* (pp. 1826–1829). Society of Exploration Geophysicists.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448–456).

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Romdhane, A., & Querendez, E. (2014). CO_2 Characterization at the Sleipner Field with Full Waveform Inversion: Application to Synthetic and Real Data. *Energy Procedia*, *63*, 4358–4365.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241).

Sinha, S., de Lima, R. P., Lin, Y., Sun, A. Y., Symons, N., Pawar, R., & Guthrie, G. (2020). Normal or abnormal? Machine learning for the leakage detection in carbon sequestration projects using pressure field data. *International Journal*

of Greenhouse Gas Control, *103*, 103189.

Wang, Z., Dilmore, R. M., & Harbert, W. (2020). Inferring CO_2 saturation from synthetic surface seismic and downhole monitoring data using machine learning for leakage detection at CO_2 sequestration sites. *International Journal of Greenhouse Gas Control*, *100*, 103115.

Williams, G., & Chadwick, R. (2021). Influence of reservoir-scale heterogeneities on the growth, evolution and migration of a CO_2 plume at the Sleipner Field, Norwegian North Sea. *International Journal of Greenhouse Gas Control*, *106*, 103260.

Wrona, T., Pan, I., Gawthorpe, R. L., & Fossen, H. (2018). Seismic facies analysis using machine learning. *Geophysics*, *83*(5), O83–O95.

Wu, X., Liang, L., Shi, Y., & Fomel, S. (2019). FaultSeg3D: Using synthetic data sets to train an end-to-end convolutional neural network for 3D seismic fault segmentation. *Geophysics*, *84*(3), IM35-IM45. Retrieved from `https://doi.org/10.1190/geo2018-0646.1` doi: 10.1190/geo2018-0646.1