CO_2 Interpretation from 4D Sleipner Seismic Images using Swin-Unet3D

Luping Qu, Kris Innanen

ABSTRACT

The interpretation of CO_2 is important for monitoring the storage status of co_2 and the risk of leakage in Carbon Capture, Utilization, and Storage (CCUS). Traditional manual interpretation of imaging dataset, while informative, is labor-intensive and often suffers from inconsistency over the extended periods of monitoring. This inconsistency largely stems from the inevitable evolution of seismic acquisition and processing technologies, as well as the subjectivity inherent in manual interpretation methods. 3D convolutional neural networks (CNNs) have seen considerable applications in object detection within seismic imaging, achieving notable success. Yet, their design constraints, specifically the limited size of convolutional kernels, have resulted in an inherent limitation in capturing long-range dependencies within the data. While Vision Transformers (ViT) excel in learning such long-distance dependencies, they are burdened by a high parameter count and struggle with local dependency information in data-scarce scenarios.

In response to these challenges, this study introduces the Swin-Unet3D model, innovatively adapted for CO_2 sequestration monitoring. This model reimagines voxel segmentation in geological imaging as a sequence-to-sequence prediction task. Its novel feature extraction sub-module is a hybrid architecture that combines the strengths of both Convolution and ViT. This parallel structure ensures comprehensive learning of both global and local dependency information within the image. The model, which is trained, validated, and tested using the Sleipner co_2 storage project's time-lapse dataset spanning from 1984 to 2010, marks an improvement in CO_2 interpretation.

INTRODUCTION

The global Carbon Capture, Utilization, and Storage (CCUS) sector has experienced a significant growth, now boasting around 40 operational commercial facilities with an annual capacity exceeding 45 million metric tons of CO2. The growth of CCUS in recent years is evident, with over 500 projects at various development stages. This surge reflects the sector's increasing importance in industrial processes, energy transformation, and power generation, playing a key role in transitioning toward sustainable, low-carbon energy systems. A milestone in the CCUS domain is the Sleipner CCS project in Norway, a pioneering endeavor in industrial-scale carbon capture and storage (Torp and Gale, 2004; Arts et al., 2004, 2008). Initiated in 1996, Sleipner has been instrumental in advancing the field, particularly in terms of long-term co_2 storage. Its innovative monitoring techniques, including time-lapse seismic and gravity field monitoring, have provided valuable insights and set industry standards (Cavanagh, 2013). The project's success in safely containing co_2 in the Utsira sandstone formation has made it a model for future CCS initiatives, demonstrating the feasibility of long-term underground co_2 storage.

Time-lapse monitoring, a critical geophysical method in CCUS projects, involves the

repeated collection of seismic data to track the behavior and containment of co_2 in geological reservoirs (**?**)lumley2021long, el2018time). This technique, essential for validating the effectiveness of co_2 sequestration, distinguishes changes in subsurface properties over time, particularly post co_2 injection. By comparing seismic data collected at different time, interpreters can visualize the growth of a co_2 plume as bright reflections in seismic images, indicative of acoustic impedance contrasts between CO2-saturated and water-saturated layers. These insights, crucial for detecting leakages and assessing storage integrity, traditionally rely on analyzing velocity variations and reflection amplitudes, necessitating expert interpretation to unravel the subtle intricacies in seismic signatures over time (Oppert et al., 2022; Chadwick et al., 2010). Accurately determining co_2 behavior in reservoirs and detecting potential leakages are complex tasks, often hampered by the limitations of traditional manual interpretation methods, which can struggle to handle the intricate variations and subtle signals present in seismic data, particularly over extended periods and across vast geological formations. Though with high-resolution Full Waveform Inversion (FWI) imaging results, co_2 interpretation requires tedious manual work and suffers from subjective errors.

The integration of Machine Learning (ML) in geophysics, particularly in co_2 interpretation, offers a transformative approach to analyzing seismic data. Compared to traditional methods, machine learning methods offer significant advantages in geophysical interpretation due to their ability to efficiently handle complex, multi-dimensional data, leading to faster, more accurate insights and adaptable, automated workflows that integrate well with technological advancements. In seismic interpretation, aachine learning methods have not only been successfully applied in seismic facies analysis, fault identification, salt body delineation, and horizon detection, demonstrating their effectiveness across a range of geophysical interpretation tasks, but they have also shown promising results in co_2 sequestration analysis. Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) represent two pioneering approaches in the realm of image analysis that have each transformed the landscape of seismic interpretation with their respective strengths (Ronneberger et al., 2015; Yuan et al., 2021). CNNs, with their inherent architecture, have been paramount in feature extraction and spatial hierarchy establishment, excelling in handling local patterns and textures inherent in medical images. The U-net architecture, for instance, has become a cornerstone in seismic data interpretation due to its effective encoder-decoder structure that meticulously captures the nuances of image features while maintaining spatial continuity. On the other hand, Vision Transformers have revolutionized the field by leveraging attention mechanisms, which facilitate the model's ability to focus on various parts of an image, thereby learning long-range dependencies that are often missed by CNNs. ViTs divide images into patches and process them as sequences, allowing for a global perspective on the image content, which is vital for understanding complex patterns that span across large regions. Despite their impressive capabilities, both CNNs and ViTs come with limitations. CNNs, by design, have a localized receptive field, which can be limiting when trying to understand the global context or long-range spatial dependencies crucial for accurate segmentation. The nature of their convolutional kernels imposes an inherent restriction on capturing long-distance relations in image data, making them less effective for tasks requiring holistic understanding of the entire scene. ViTs, while capable of modeling long-distance interactions, often require substantial computational resources and can be inefficient with smaller datasets. Their performance, heavily reliant on large amounts of

data for training, can lead to challenges in medical segmentation tasks where fine-grained details are critical. Recognizing the complementary nature of CNNs and ViTs, the introduction of a model, named Swin-Unet3d, that synergizes the advantages of both holds significant promise. This hybrid model aims to address the limitations of both approaches by integrating the local feature extraction capabilities of CNNs with the global contextual understanding of ViTs. In the context of co_2 interpretation in geophysical applications, this model adapts the principles that have driven success in medical segmentation to address the unique challenges presented by subsurface imagingc(Wu et al., 2019; Li and Li, 2021). co_2 interpretation requires a nuanced approach that can discern subtle variations indicative of co_2 existence within geological formations. Swin-Unet3D, taking cues from successful applications in medical imaging, which adeptly combines the granular detail-oriented approach of CNNs with the broad, contextualized perspective of ViTs, offers a promising avenue for enhancing the precision and reliability of co_2 interpretation.

In our research, we introduce a novel neural network approach for delineating co_2 distribution from time-lapse seismic dataset using both baseline and monitor imaging processing results. Our method is anchored by a streamlined version of the Swin-Unet3D architecture, a model that has already demonstrated its effectiveness in 3D medical image segmentation. The model used in this study integrates the Swin Transformer Block3D for feature extraction in 3D medical images, paralleled with a traditional Convolution3D approach. This dual-structured model capitalizes on the strengths of both the Swin Transformer and CNN architectures: the Swin Block3D module excels at identifying long-range dependencies within images, while the Conv Block3D focuses on short-range, localized information. In the decoding phase, features extracted by these two modules are meticulously fused, leveraging the broad contextual insights of the Swin Transformer with the detailed local understanding of the CNN. This fusion results in a more holistic interpretation of the image data. The model's efficacy is particularly evident in complex tasks like image segmentation, where the combination of Vision Transformer and Convolutional structures compensates for each other's limitations, leading to enhanced accuracy and robustness in predictions. We adapt it for geophysical analysis by training on a publicly available dataset from Sleipner carbon capture and storage (CCS) site. The trained neural network is additionally employed to forecast the co_2 dispersion for other years which do not have interpretation results, utilizing baseline and monitoring seismic imaging as inputs. Leveraging a single A100 GPU with 24 GB of memory, we complete training and validation in roughly one hour. Upon comparing the outcomes with those from a CNN-based model, we observed that our network exhibits superior accuracy in predicting the boundaries, highlighting its enhanced performance in delineating edges and contours.

METHODS

Swin-Unet3D Architecture

Swin-Unet3D, which is tailored to capture the intricate spatial distribution of CO2, is a combination of an encoder, a jump connection, and a decoder, where the encoder harnesses the Swin Transformer Block3D to interpret long-range dependencies, essential for understanding broader image contexts, as shown in FIG 1 (Cai et al., 2023). Concurrently, the Conv Block3D within the encoder specializes in discerning the localized, short-range infor-

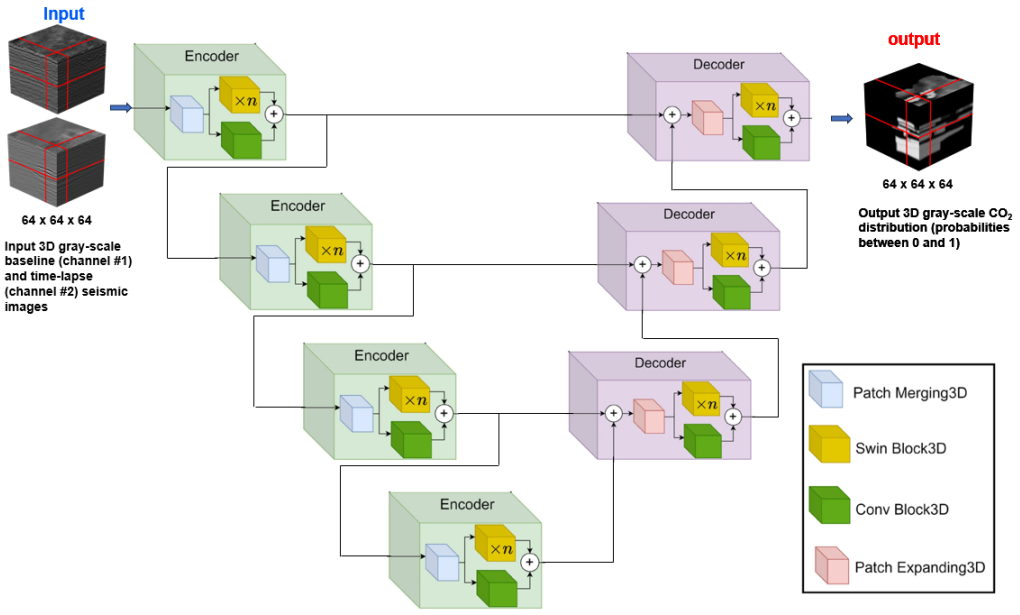

FIG. 1. The architecture of the Swin-Unet3D.

mation, crucial for detailed feature extraction. The Swin-Unet3D architecture facilitates a seamless feature analysis by dividing 3D seismic images into multiple voxel blocks, subsequently flattening each voxel block into one-dimensional vectors, which are then processed through the network's encoder and decoder pathways. The Patch Merging3D module assists in downsampling, while the Patch Expanding3D module is pivotal in the decoder stage, upscaling the feature vectors to reconstruct the image's spatial resolution.

The next step is to encode the image: we divide the height H, width W, and depth D by 4 to get the number of tokens, which are vectors of length N. For a standard setup, N is usually 96, consistent with Swin Transformer guidelines. These tokens are then processed by Conv Blocks3D and Swin Blocks3D within the model, which extract features from the image data. These features pass through the Patch Expanding module and multiple decoders to rebuild the spatial details. After several decoding stages, we can recover the original image's spatial resolution. The final image classification or segmentation can then be done by applying a simple linear transformation to the last layer of the extracted features.

Each layer in this process includes a multi-head attention mechanism, which varies in number at different stages of the model. For instance, in Swin-Unet, the Encoders (from Encoder1 to Encoder5) utilize varying counts of these attention mechanisms, specifically 3, 6, 12, 24 respectively. Likewise, the Decoders (from Decoder4 to Decoder12) also use different counts, in this case, 12, 6, 3 respectively. Moreover, the number of Swin Blocks3D in each Encoder varies, with Encoder1 to Encoder5 containing 2, 2, 6, 2 respectively, and Decoder4 to Decoder12 having 6, 6, 2 respectively.

The introduction of the submodule of the network including the detailed introduction of swin Block3D, Conv Block3D, Patch Merging3D, Patch Expand3D are illustrate in Cai et al. (2023)'s work.

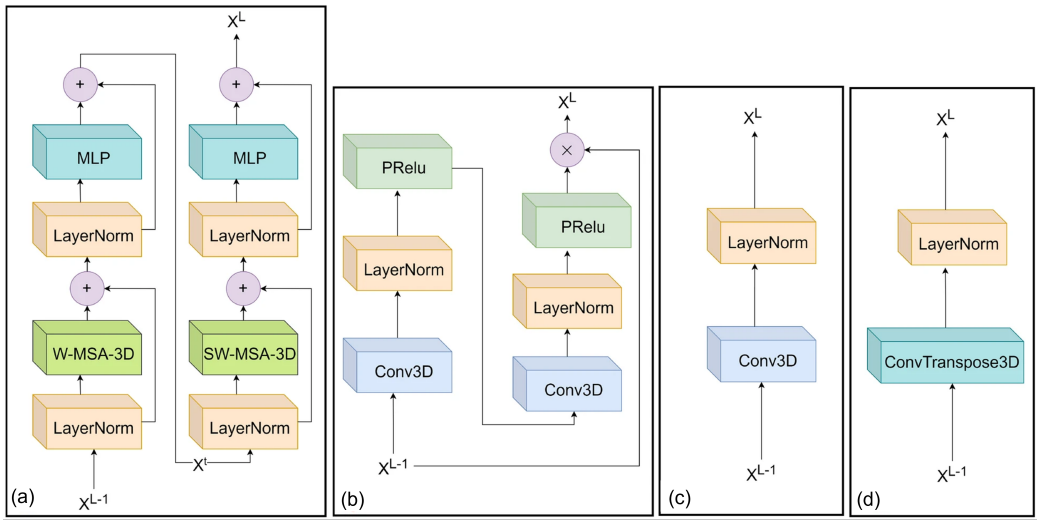

FIG. 2. Overview of the structure of some sub-modules: a swin Block3D, b Conv Block3D, c Patch Merging3D, d Patch Expand3D ().

Encoders and decoders

The architecture of both the Encoder and Decoder components is depicted in FIG 1. Each Encoder is comprised of a Patch Merging3D sub-module, a Conv Blocks3D sub-module, and multiple Swin Blocks3D sub-modules. The Patch Merging3D sub-module generates an interim feature image by downsampling either the input image or the output from the preceding Encoder. The Conv Blocks3D and Swin Blocks3D sub-modules concurrently process the feature image, where the former handles short-distance dependencies and the latter focuses on capturing long-distance dependencies within the image. An addition of matrices from the Conv Blocks3D and Swin Blocks3D outputs results in the final feature image from the Encoder.

The Decoder's structure mirrors the Encoder's, with the key distinction being its dual input configuration. The first input comes from the feature image produced by the corresponding Encoder of the same hierarchical level, and the second input is sourced from the feature image output of the preceding Decoder. This dual input scheme in the Decoder is designed primarily to embed residual connections, which are crucial for preventing the diminishing gradient issue during the backward propagation phase, as noted in the literature.

Data preparation

The Sleipner project dataset provides a detailed representation of a geological storage formation. This dataset is unique as it includes time-lapse seismic records capturing the progression of co_2 injection and its dispersal into distinct layers of the geological formation.

At the core of this dataset are nine layers of sandstone, named Utsira L1 through Utsira L9 (FIG. 3). These layers are central to the project as they have been identified as the primary storage sites for the injected CO2. Each layer is a potential reservoir, with L1 being the lowest and L9 the uppermost. Separating these sandstone layers are thin, low-permeability barriers composed of shales and mudstones. These horizontal barriers, known

as "intrashales," are approximately 1 meter thick and serve to compartmentalize the storage formation into discrete sections. An exception to this pattern is found between layers L8 and L9, where a substantially thicker shale layer, approximately 7 meters thick, known as the "Thick Shale Unit," creates a more pronounced separation. The project has evolved significantly since its initial benchmark model in 2011, which only covered the topmost layer, L9. The 2019 Benchmark Model, however, marks a significant advancement, offering the first complete 3D model of the storage site encompassing all nine layers. This comprehensive model allows for a full-scale analysis of the subsurface co_2 distribution across all strata. Above the sandstone formations lies a protective caprock, a 50-meter thick shale layer incorporated into the dataset as a uniform zone. While the actual caprock is thicker, this simplified representation is sufficient for most modeling purposes. The caprock's role is critical as it acts as a seal to ensure the co_2 remains contained within the sandstone reservoirs below. The dataset presents the co_2 plumes' spatial distribution within the sandstone layers, with data points starting from the year 2010. These plumes in FIG. 3 are visual representations of the co_2 that has migrated into the formation post-injection, creating separate "plumes" in each sandstone layer.

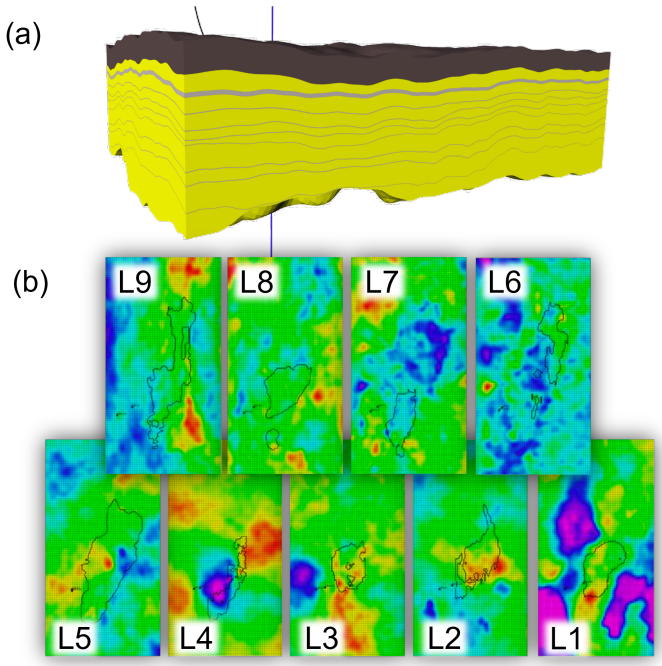

FIG. 3. The geology information of sleipner site and co_2 boundary at each horizon.

The co_2 existence interpretation can be treated as either an image segmentation or object detection problem. This study aims to delineate the CO2-saturated regions with well-defined boundaries by assigning a value of one to each pixel representing the target and zero to all others, as shown in FIG **??**. The input data consists of two channels. These channels represent the inversion imaging results from seismic data collected from the years 1994 and 2010, respectively. Each 3D cube has dimensions of 64 $\times$ 64 $\times$ 64, covering a volume of 1.6 km $\times$ 1.6 km $\times$ 512 ms in real distance, situated within the overall data volume of 3.35 km $\times$ 7.0 km $\times$ 2,000 ms. Normalization is conducted separately for the input cubes of the baseline and monitor images through subtracting the mean value and dividing by the standard deviation for each image cube, as the same with Li and Li (2021)'s work.

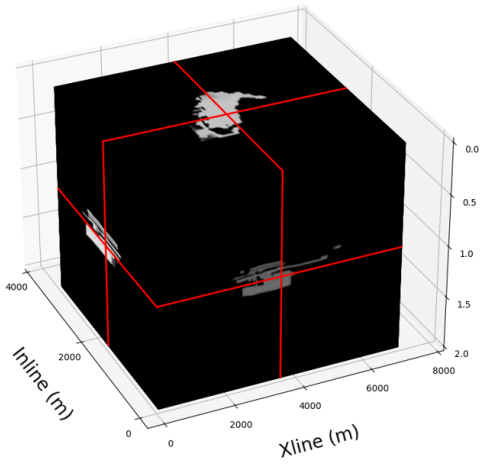

FIG. 4. The geology information of sleipner site and co_2 boundary at each horizon.

The Sleipner 4D seismic dataset, made available by Equinor in 2020, is a comprehensive collection of both baseline and time-lapse seismic recordings that have undergone time-lapse processing across various years. The initial baseline seismic data were collected in 1994, which predates the commencement of co_2 injection in 1996. Subsequent time-lapse seismic surveys were systematically conducted, yielding datasets from the years 1999, 2001, 2004, 2006, 2008, and 2010. These datasets are designated using a unique nomenclature where 'xxpyy' indicates data acquired in the year 'xx' and processed in the year 'yy'. Corresponding to each time-lapse set, there is a reprocessed baseline dataset labeled as '94pyy', which aligns with the year of the baseline data acquisition. The sole exception within this series is the 10p11 dataset, which has been subjected to image processing without an associated reprocessed baseline dataset. As our focus is solely on the near-offset images, only the near-offset images are used in this study. This subset has been determined to be the most pertinent for interpreting co_2 movements within this specific research framework.

FIG 5 shows two baseline imaging of the data collected from 1994 and processed in 2001 and 2010 respectively. These images are essential for understanding the geological conditions before the start of co_2 injection, with distinct horizontal layering in the subsurface. FIG 6 shows the monitor imaging of the data collected from 2010 and processed in 2010 and 2011 respectively. These images capture the changes in the subsurface following co_2 injection, with the co_2 plumes being visible as anomalies within the sandstone layers. FIG 7 displays the frequency spectra. The average frequency spectra of the 94p10 and 10p10 datasets are closely matched, with only minor variations observed when compared to the 94p01 dataset. However, the 10p11 dataset stands out with a noticeable difference in its bandwidth due to the lack of time-lapse processing. These variations between the datasets reflect the processing inconsistencies between the baseline and time-lapse images, which are critical for increasing the resilience of the neural network to discrepancies caused by processing in image space.

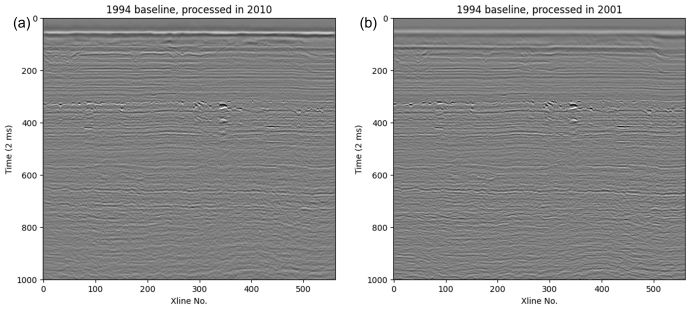

FIG. 5. The inline assemble at 1,625 m for baseline processing images from (a) 94p01 and (b) 94p10.

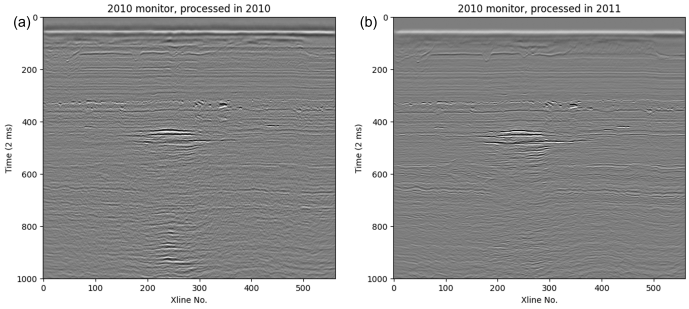

FIG. 6. The inline assemble at 1,625 m for monitor processing images from (a) 10p10 and (b) 10p11.

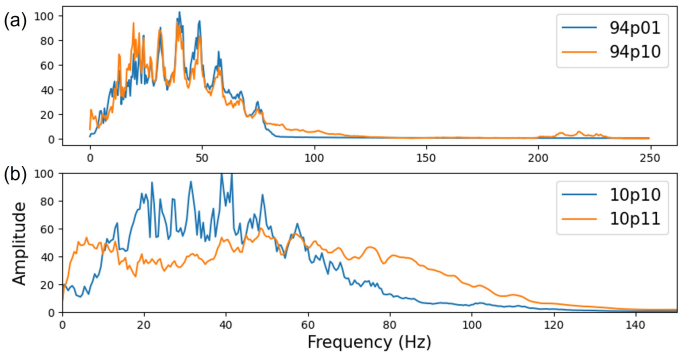

FIG. 7. Amplitude spectra of (a) 94p01 and 94p10, (b) 10p10 and 10p11. The data set amplitude spectra differences are caused by both artificial effects, such as acquisition and processing differences.

Experiment on the time-lapse dataset of Sleipner

Training and validation

In our study, we selected 3D cubes at random from one baseline, one time-lapse, and the corresponding label volumes. Initially, we extracted 1000 cubes with centers distributed randomly throughout the 3D volume, as shown in FIG 8. Several samples of the sampled cubes in small size are shown in FIG 9. In FIG 9, Each cube is paired with four different sets of baseline and time-lapse inputs, yielding a total of 4000 samples. However, we noted that approximately 60% of these samples contained no labeled co_2 plume, being entirely zero-valued in their labels, which considerably delayed the model's learning process. These

samples were then randomly split into training and validation sets, containing 3500 and 500 samples, respectively. The configuration of the training and validation sets is introduced in Table 1.

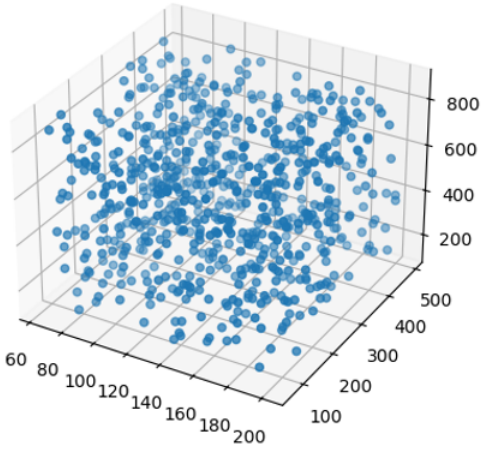

FIG. 8. The center points of the sampled cubes generated randomly.

Table 1. Training setup

Loss function	**Optimizer**	**Learning rate**
Binary cross-entropy (BCE)	Adam	0.0005
Batch size	**software**	**Hardware**
30	Pytorch 2.1.0	GPU-A100 (32G)

FIG 10 illustrates the progression of training and validation losses as the number of epochs increases. FIG 10a is the loss curve of a CNN training for reference, and FIG 10b is the loss curve of Swin-Unet3D training. Totally 200 epochs are chose through checking the validation loss. The total training takes approximately 1 hour using one node on GPU-A100. Both the training and validation losses exhibit a significant reduction, shrinking by about three orders of magnitude, which indicates that the model is learning and improving its predictions over time. The two curves are close to each other, which suggests that the model is generalizing well rather than overfitting to the training data. There are small spikes in the validation loss, which are normal in training dynamics. They can occur due to various reasons, such as the learning rate, the nature of the mini-batches, or the model encountering new patterns it previously hadn't learned from the training data. FIG 10b indicates a successful training process where the model has learned to generalize the patterns from the training data effectively, as evidenced by the convergence and stabilization of the training and validation loss.

FIG **??** depicts the neural network's predictions using CNN and Swin-Unet3D for two samples from the training dataset. The neural network's (NN) predictions generated using the model trained by CNN and Swin-Unet3D both show a high degree of agreement with the actual labels, while the CNN predictions are less sharp in resolution. In contrast, predictions made by Swin-Unet3D demonstrate more defined boundaries and a greater resolution

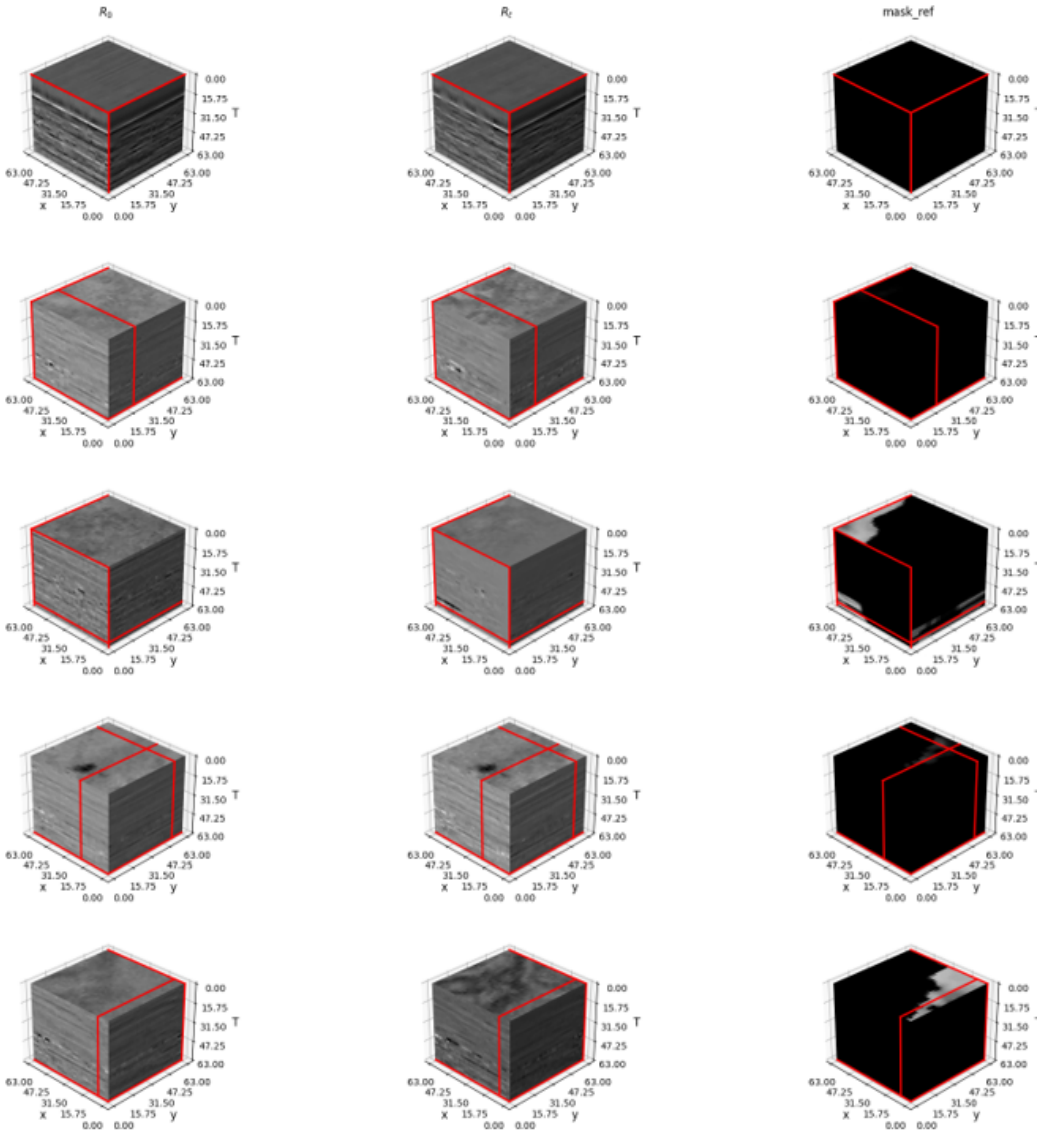

FIG. 9. Examples of the sampled input and label data. The first column shows the sampled baseline data, the middle column shows the sampled monitor data, and the right column shows the sampled label data

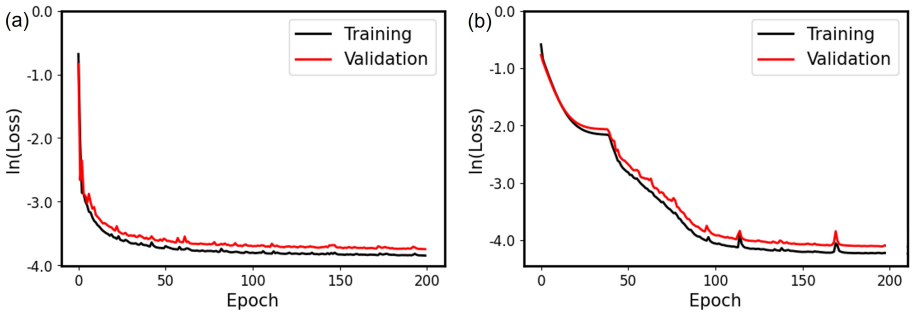

FIG. 10. Trained NN predictions for two samples from the training data set. The first two rows are the baseline data and monitor data for these two cube samples. The third row are the label data. The fourth row are the NN predictions generated using CNN, and the last row are the NN predictions generated using Swin-Unet3D.

in comparison to those from CNN, showcasing the advanced capability of Swin-Unet3D in capturing detailed features. Similar results are shown in FIG 12.

RESULT AND DISCUSSION

In our approach, we finely segmented the entire 3D volume into small, regularly overlapping cubes as input for our neural network. We then synthesized the comprehensive 3D CO2 distribution through weighted summation.FIG 12 shows the whole target volumn predictions for both CNN and Swin-Unet3D trained network. When analyzing the prediction results, the similarities between the label data and the NN's predictions are strikingly close, with minimal discernible differences.

CONCLUSIONS

In this study, we have harnessed the capabilities of the Swin-Unet3D model to interpret the 3D distribution of CO_2 from a series of time-lapse seismic datasets. Our findings demonstrate that the Swin-Unet3D model, with its innovative architecture that synergizes the strengths of CNNs and Vision Transformers, offers a more intricate and detailed characterization of CO_2 plumes compared to CNN approaches. The Swin-Unet3D's ability to capture long-range dependencies and intricate details significantly enhances the accuracy of CO_2 boundary delineation within the seismic images. The application of this model to the Sleipner project's dataset has not only expedited the interpretation process but has also provided a level of consistency and detail that surpasses conventional manual interpretation methods. The improved resolution and sharpness of the boundaries in the model's predictions underscore its potential to transform CO_2 monitoring practices in Carbon Capture, Utilization, and Storage projects. Our results indicate that incorporating advanced machine learning models like Swin-Unet3D into the workflow can lead to more reliable and efficient monitoring of CO_2 storage sites. The potential for these technologies to contribute to safer, more effective CCUS practices is substantial and holds promise for broader application in the geosciences field.

Future work will aim to refine these methods further, potentially integrating real-time data processing to provide immediate insights into CO_2 storage dynamics. As the field progresses, we anticipate that machine learning will play an increasingly pivotal role in

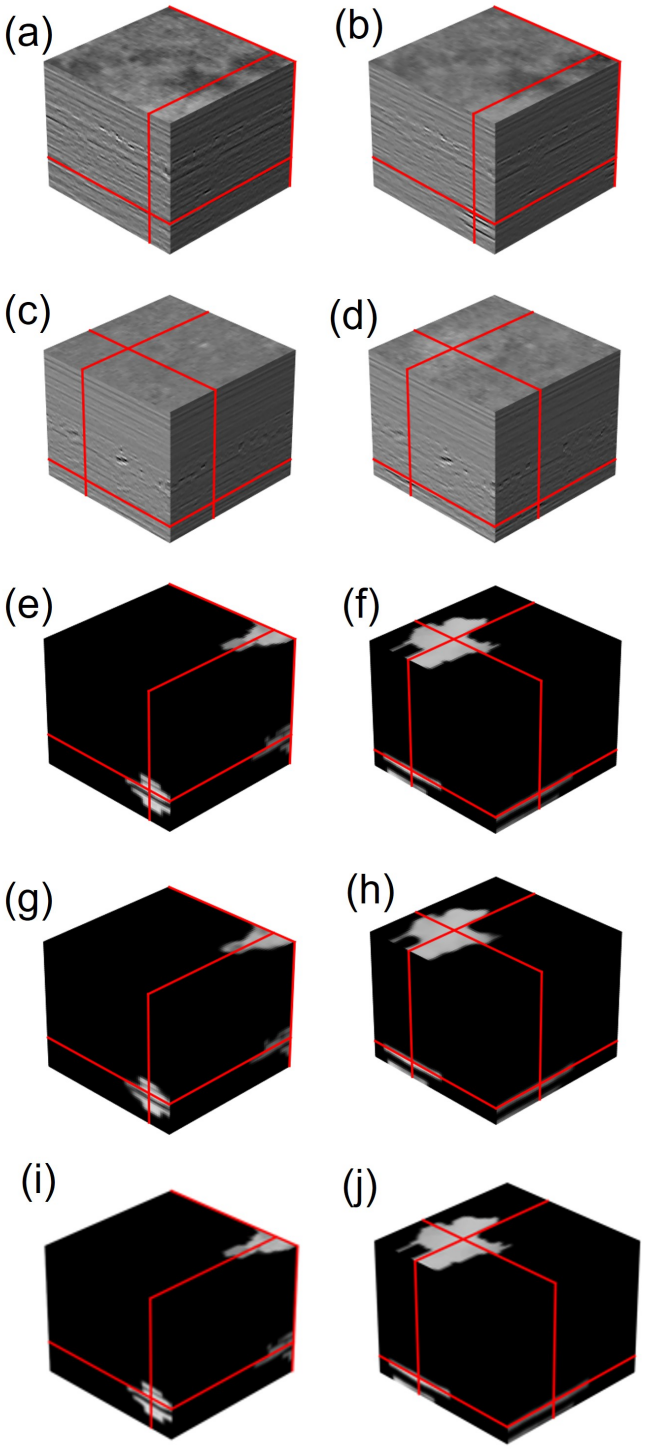

FIG. 11. Trained NN predictions for two samples from the training data set. The first two rows are the baseline data and monitor data for these two cube samples. The third row are the label data. The fourth row are the NN predictions generated using CNN, and the last row are the NN predictions generated using Swin-Unet3D.

advancing the safety, efficiency, and reliability of CO_2 sequestration efforts globally.

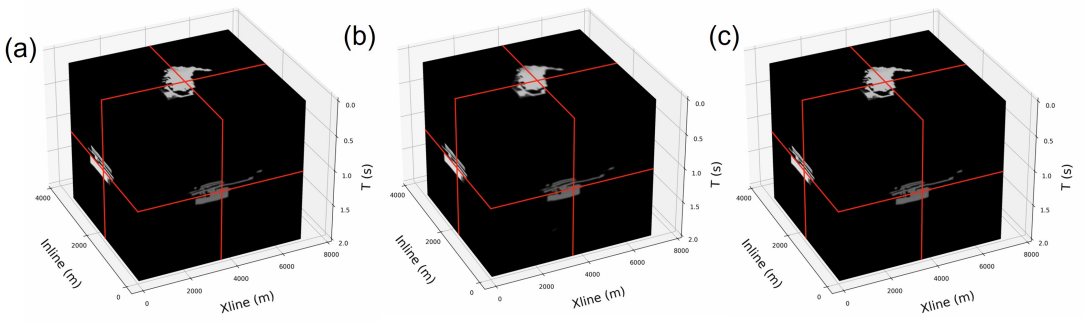

FIG. 12. Trained NN predictions for the whole volumn, (a) is the ground-truth label data, (b) is the predicted result of the model train by CNN, (c) is the predicted result of the model trained by Swin-Unet3D.

ACKNOWLEDGEMENTS

The sponsors of CREWES are gratefully thanked for continued support. This work was funded by CREWES industrial sponsors, NSERC (Natural Science and Engineering Research Council of Canada) through the grant CRDPJ 543578-19.

REFERENCES

Arts, R., Chadwick, A., Eiken, O., Thibeau, S., and Nooner, S., 2008, Ten years' experience of monitoring co2 injection in the utsira sand at sleipner, offshore norway: First Break, **26**, No. 1, 65–72.

Arts, R., Eiken, O., Chadwick, A., Zweigel, P., Van der Meer, L., and Zinszner, B., 2004, Monitoring of co2 injected at sleipner using time-lapse seismic data: Energy, **29**, No. 9-10, 1383–1392.

Cai, Y., Long, Y., Han, Z., Liu, M., Zheng, Y., Yang, W., and Chen, L., 2023, Swin unet3d: a three-dimensional medical image segmentation network combining vision transformer and convolution: BMC Medical Informatics and Decision Making, **23**, No. 1, 33.

Cavanagh, A., 2013, Benchmark calibration and prediction of the sleipner co2 plume from 2006 to 2012: Energy Procedia, **37**, 3529–3545.

Chadwick, A., Williams, G., Delepine, N., Clochard, V., Labat, K., Sturton, S., Buddensiek, M.-L., Dillen, M., Nickel, M., Lima, A. L. et al., 2010, Quantitative analysis of time-lapse seismic monitoring data at the sleipner co 2 storage operation: The Leading Edge, **29**, No. 2, 170–177.

Li, B., and Li, Y. E., 2021, Neural network-based co2 interpretation from 4d sleipner seismic images: Journal of Geophysical Research: Solid Earth, **126**, No. 12, e2021JB022,524.

Oppert, S., Adachi, J., Thornton, D., and Royle, A., 2022, Monitoring technology to enable characterization of ccus reservoirs, *in* SEG International Exposition and Annual Meeting, SEG, D011S199R001.

Ronneberger, O., Fischer, P., and Brox, T., 2015, U-net: Convolutional networks for biomedical image segmentation, *in* Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 234–241.

Torp, T. A., and Gale, J., 2004, Demonstrating storage of co2 in geological reservoirs: The sleipner and sacs projects: Energy, **29**, No. 9-10, 1361–1369.

Wu, X., Liang, L., Shi, Y., and Fomel, S., 2019, Faultseg3d: Using synthetic data sets to train an end-to-end convolutional neural network for 3d seismic fault segmentation: Geophysics, **84**, No. 3, IM35–IM45.

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.-H., Tay, F. E., Feng, J., and Yan, S., 2021, Tokens-to-token vit: Training vision transformers from scratch on imagenet, *in* Proceedings of the IEEE/CVF international conference on computer vision, 558–567.