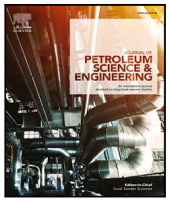
Journal of Petroleum Science and Engineering 215 (2022) 110648

Contents lists available at ScienceDirect

Journal of Petroleum Science and Engineering

journal homepage: www.elsevier.com/locate/petrol



# Size-invariant 3D generation from a single 2D rock image

Johan Phan [a,b,*], Leonardo Ruspini [b,c], Gabriel Kiss [b], Frank Lindseth [b]

[a] NTNU, Høgskoleringen 1, Trondheim, 7491, Norway
[b] Petricore Norway, Stiklestadveien 1, Trondheim, 7041, Norway
[c] Q-free, Strindfjordvegen 1, Trondheim, 7053, Norway

ARTICLE INFO

*Keywords:*
Porous media
Digital rock
2D to 3D
Reconstruction
Deep learning

ABSTRACT

The characterization of 3D structures in porous media is crucial for predicting physical properties in many industries, such as CO2 capture and storage, hydrology, oil & gas. In contrast to the expensive and time-consuming acquisition of 3D images, 2D imaging can provide cheap and fast data. However, the reconstruction of a 3D image from a single 2D image is a complex non-deterministic inverse problem. Several statistical and deep learning-based algorithms have been introduced in the past, however, most of them fail to generalize structures and textures for different types of rocks, in addition to being time-consuming and only able to generate relatively small images ($300^3$ voxels cube).

In this work, we propose a size-invariant multi-step 3D generation workflow from a single 2D image using a combination of Vector-Quantized Variational AutoEncoder(VQ-VAE), size-invariant Generative Adversarial Networks(GAN), and Image Transformer. The proposed workflow tackles several major challenges in the generation of 3D images since it is designed to not only satisfy the large size constraint ($>1000^3$ voxels cube) but also to generate statistically representative pore structures. The combination of these different generative techniques allows us to overcome the scalability, stability, and complexity associated with GAN approaches.

We trained the proposed workflow using several types of rocks with different physical properties, sizes, and resolutions. To validate our methodology, we have generated several large-size 3D rock images and compare them to real 3D images in terms of physical properties (porosity, permeability, and Euler characteristic).

## 1. Introduction

Characterization of porous media properties such as transport properties, storage capacity, and capillary-trapping, is an essential step in many important applications including CO2 capture and storage, underground water management, oil & gas reservoir management, and material sciences. The derivation of these properties from three-dimensional computed tomography(CT) images is a disruptive technology that can fundamentally change the way porous media are characterized. However, the process of acquiring CT images is expensive, time-consuming, and limited to a certain resolution range given by the scanner. The ability to generate realistic 3D micro-structures from 2D images can significantly extend the technology's capabilities by using, for example, cheaper and higher resolution SEM or thin section images.

Several mathematical methods have been presented in the literature to generate a 3D image from a single 2D image; among them, process-based modeling (Bakke and Øren, 1997; Øren and Bakke, 2002) and stochastic-based methods (Strebelle, 2002; Tahmasebi et al., 2012, 2014) have been used and developed over the years. Process-based modeling simulates the rock formation processes, which in many cases can be extremely complex or even unknown. On the other hand, stochastic-based methods use statistical information from the 2D image to stochastically populate a 3D volume. However, these processes are time-consuming and often result in quite homogeneous porous structures (Pant, 2016; Okabe and Blunt, 2004; Čapek et al., 2009). In addition, both process-based and stochastic-based methods require a high degree of iterative interaction with expert users, e.g. geologists, to input a significant number of parameters which in general are adjusted by trial and error.

In recent years, several deep learning-based methods, primarily GAN, have been applied for 3D porous media image generation (Mosser et al., 2017, 2018; Feng et al., 2019, 2020). Although the proposed deep learning-based methods promise a faster and fully-automatic way to generate 3D images, these studies have been limited to a single type of rock for each model and only managed to generate small, fixed-size images ($64^3$ to $256^3$ voxels cube). As described in Bruns et al. (2017), large enough 3D rock images are required to obtain representative physical properties, especially for complex heterogeneous rocks where the representative size can be several thousand voxels in each

* Corresponding author at: NTNU, Høgskoleringen 1, Trondheim, 7491, Norway.
*E-mail addresses:* johan.phan@ntnu.no (J. Phan), leonardo.ruspini@petricore.com (L. Ruspini).

https://doi.org/10.1016/j.petrol.2022.110648
Received 16 March 2022; Received in revised form 28 April 2022; Accepted 13 May 2022
Available online 21 May 2022

direction (Ruspini et al., 2021). Consequently, a significant increase in generation size needs to be achieved before deep learning-based methods can be used in real-world applications.

Therefore, the main goal of this work is to develop an automatic workflow for 3D generation from a single 2D rock image. The design requirements for this new workflow are:

- It should generate statistically representative 3D images from a single 2D image.
- It should be capable of generating 3D images of any size, and it should work with input images of any resolution and size.
- Since rock samples from different places are unique due to their formation processes; we would expect images of new types of rock to be periodically added to our training dataset. Therefore, our models should be stable and robust during training so that a continuous training pipeline can be applied to incorporate new rock types periodically. This requirement poses a limitation to GAN-based approaches, well known for being unstable during training with changing datasets due to catastrophic forgetting and mode collapse (Thanh-Tung and Tran, 2020).
- The total generation time for large images (>$2000^3$ voxels cube) should be limited to not more than a day.

## 2. Related works

### *3D Rock image generation*

Generating a statistical representative 3D rock image from a single 2D image is a long-standing problem in the field of digital rock analysis. Traditionally, process-based modeling (Bakke and Øren, 1997; Øren and Bakke, 2002) and stochastic methods (Adler et al., 1990; Strebelle, 2002; Blair et al., 1996; Tahmasebi et al., 2012, 2014) have been used for 3D reconstruction using indirect information from 2D images.

**Process-based modeling** is based on simulating the processes involved in rock formation (e.g., sedimentation, compaction, and diagenesis). However, this method limits its use to relatively simple and homogeneous rock types, e.g., Bentheimer sandstone. Even if the simulated processes continue evolving for more complex rock types such as carbonates (Ruspini et al., 2021), they fail to describe multi-scale structures and variations intrinsically associated with a large part of reservoir rocks.

The **stochastic methods**, on the other hand, construct 3D structures using spatial statistical constraints extracted from 2D images. These methods are often time-consuming (tens of hours for $300^3$ voxels cubes) and require simplifications of rock structures to work properly, i.e., in general, they produce unrealistic-looking 3D images.

Recently, **deep learning-based methods** have gained popularity due to the rise of deep learning technology. With sufficient data and computational power, deep learning-based methods can theoretically be trained to generalize the information in a single 2D image to reconstruct a corresponding 3D image. Among the deep learning methods, GAN (Goodfellow et al., 2014) has been favored in most recent works. The first proposed GAN models only worked for a single rock type and did not produce satisfactory results for all the cases (Mosser et al., 2017, 2018). In later works, Conditional-GAN models were used to generate a 3D rock image with a 2D image as input (Volkhonskiy et al., 2019; Valsecchi et al., 2020; Zhao et al., 2021; Coiffier et al., 2020). A hybrid model combining GAN and Variational autoencoder (VAE) (Zhang et al., 2021) has also been proposed for a more stable training process. All the mentioned works were limited to relatively small size image generation (between $64^3$ and $256^3$ voxels cube). Even using a size invariant GAN to generate larger images is possible, training on a large dataset with multiple types of rocks would still be an extremely challenging problem. This problem is due to the implicit nature of GAN-based generation methods, where a 3D image is generated from a latent vector sampled from a random distribution. In most cases, this randomly generated latent vector also defines the structure of the generated image. This poses a fundamental problem of using a GAN since instead of gradually propagating the structural information from the 2D to the 3D image, GAN uses the latent vector to generate the structure of the 3D image directly and then adjust the results to fit the 2D input image. Moreover, a continuous distribution latent space is unsuitable for describing a dataset consisting in different types of rock, where each type has a distinctive structure.

### *Autoregressive models for image synthesis*

GAN models have traditionally dominated the landscape of deep learning-based image synthesis; however, they are notorious for being difficult to train due to the mode collapse problem (Thanh-Tung and Tran, 2020). On the other hand, autoregressive models with tractable likelihood such as PixelCNN (Oord et al., 2016), VQ-VAE-2 (Razavi et al., 2019), Image Transformer (Parmar et al., 2018), and ImageGPT (Chen et al., 2020) are simpler to train and better to capture the diversity in the data distribution. In addition to 2D image synthesis, autoregressive models have also been successfully applied in video prediction from a single image where VQ-VAE and GPT were combined to achieve state-of-the-art results (Yan et al., 2021). The main drawback of using autoregressive models is that they have a slow inference time due to the need to sequential sample pixel-by-pixel, and therefore they struggle with scalability, especially for large image generation.

In the next section, we propose a method that uses VQ-VAE to massively compress large images into compact discrete representation vectors even at the cost of high information loss so that an autoregressive model can be applied more effectively. To compensate for the information loss, we also employ a GAN model in our method.

## 3. Method

Fig. 1 illustrates our new approach to generate large-scale 3D images from a single 2D image. The proposed method consists of 3 main components:

**1. VQ-VAE** (Oord et al., 2017; Razavi et al., 2019): VQ-VAE is an encoder–decoder based model that maps an input image into a discrete representation vector. In this work, we use VQ-VAE to compress the input 3D image by $16^3$ times using a code-book of 128 indices. This high compression level helps extract the structural information from the 3D image such as pore distribution and grain structure from a rock image into an information-dense discrete vector.

**2. Image Transformer** (Parmar et al., 2018): Image Transformer is an autoregressive generation technique that uses self-attention mechanisms to model the distribution of image contents with tractable likelihood. Since the memory consumption of transformer scales quadratically with the input length (Gupta et al., 2021), working directly with these 3D images would not even cover a single grain. Therefore to maximize the receptive fields for tracking spatial dependencies, we train our transformer on the compressed vector produced by the VQ-VAE. In order to generate large images, we use 3D *Local Attention* with a maximum receptive field of 16 × 16 × 10 voxels to query voxel by voxel in a raster-scan order. This approach is similar to how pixelCNN was used in the VQ-VAE-2 (Razavi et al., 2019), but in a 3D context where the PixelCNN is replaced with a transformer. The transformer used in this work uses the GPT-3 architecture, see Brown et al. (2020), but has only a vocabulary size of 128 words, 16 attention heads, 16 layers, and a total of 50 million trainable parameters.

**3. Size invariant GAN**: We loosely based the architecture of our GAN on the residual-based architecture proposed in StyleGan-v2 (Karras et al., 2020) for effective gradient passing when training. The use of GAN, in this case, serves a similar role as a super-resolution model, i.e., populate texture and high-resolution details to the structures generated in the previous steps.

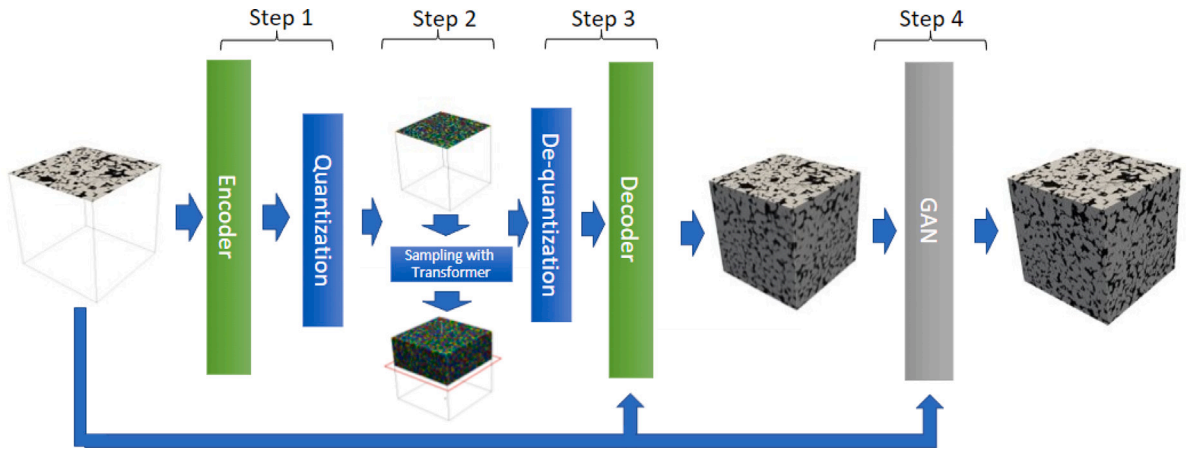



**Fig. 1. 3D generation workflow from a single 2D image** – Our generation workflow consists of 4 steps: (1) Encode/quantize to compress the image into a compact and discrete representation vector of $16^3$ times smaller size; (2) Iteratively sample the next data-point of the quantized vector with a transformer until the desired size is met; (3) Decompress and de-quantize this generated vector; (4) Regain the detailed textures with GAN. In addition, we also give the decoder and the GAN the 2D input as additional information.

*Method description*

First, the 2D input image is passed through the encoder to get the first layer of the quantized vector (Fig. 1 step 1). Then, we complete the quantized vector from the first layer by iteratively passing it through the transformer to get the likelihood distribution of the next data point and sample it with a top $k$ sampling strategy ($k = 10$) (Fan et al., 2018) (Fig. 1 step 2). Using a sampling strategy (e.g., top-$k$, nucleolus sampling) is imperative to prevent cascade degeneration since our generation target is often significantly longer than the model receptive field of view (Holtzman et al., 2019).

This workflow allows the transformer to focus on the global structures while the GAN generates the detailed texture (Fig. 1 step 4). The approach is well suited for the nature of the problem since the structural features of a rock image, e.g., grains packing, are strictly bounded by the statistical and physical properties of the rock type. On the other hand, detailed features such as the surface texture of each grain are local and less complex to generate, thus are suited for GAN. Another motivation to combine a transformer with a VQ-VAE and GAN is to maximize the receptive field to capture the spatial image context. Although a $16^3$ voxels image is a small volume, it has a sequence length of 4096 words when flattened, which is longer than most Natural Language Generation (NLG) models input length. Therefore, the use of VQ-VAE allows the transformer to have a spatial context window equivalent to $256^3$ voxels cube in terms of image structure even with just a $16^3$ voxels input.

*Size-invariant model*

Given that rock images, in general, are not restricted to a fixed size or dimension ratio, see Table 1, it is important to create a workflow that is independent of the input size, i.e., size-invariant. Building a model entirely with local operations such as convolution, pooling, and quantization, allows us to theoretically generate images of any size as long as the input image size is divisible by the size of the receptive field, e.g. $64^3$ voxels cube in our case. However, processing 3D images with CNN is very memory demanding, e.g. our model requires over 20 GB of GPU memory to generate/process a $196^3$ voxel cube image. Since most real-world scenarios require images larger than $1024^3$ voxels to cover the representative volume, it is imperative to split the input image into smaller patches and process each patch separately. However, naively putting the generated image patches together into a large image would result in an unnatural transition between each patch due to the border effect caused by zero-padding, as shown in Fig. 2(a). Since zero-padding provides CNNs with many important benefits, such as allowing CNNs to encode position information and keeping the spatial size constant


**Fig. 2. Output crops from the output of VQ-VAE** – Our merging strategy allows a smooth transition when combining patches into a large image.

after each layer (Islam et al., 2021; Kayhan and van Gemert, 2020), one cannot simply refrain from using zero padding to avoid the border problem. One way to deal with this problem is to use a merging strategy. In Fig. 3 we illustrate our merging strategy where we process the image patch-by-patch in a similar way to a sequential convolutional operation. To generate a large image with a smooth transition, as shown in Fig. 2(b), we introduce the concept of **base-cube** in both training and inference, Fig. 4. The size of the **Base-cube** is defined as the maximum receptive field, i.e. the size of the region in the input that produces the feature, of our model.

When generating an image using base-cubes, we can eliminate the border effect by simply discarding the cubes affected by the padding, as shown in gray in Fig. 3. Since our objective is to merge the base-cubes into a larger image, we need to train the model to learn to combine the base-cubes. Therefore, instead of training with a single base-cube, we vary the input size of the training data so that each batch consists of multiple cubes arranged in a random order, as shown in Fig. 4

*Training data*

In this work, we use porosity images to train and evaluate our models. To build a porosity image, two images are acquired by scanning the rock sample in dry and brine–saturated states. Then a porosity image is constructed as the difference between the two images, with an attenuation correction as described in Arns et al. (2003), Golab et al. (2010). Our training dataset contains 9 porosity images of different types of rocks (3 sandstones and 6 carbonates) with different resolutions. They were acquired in different X-ray scanners, using different

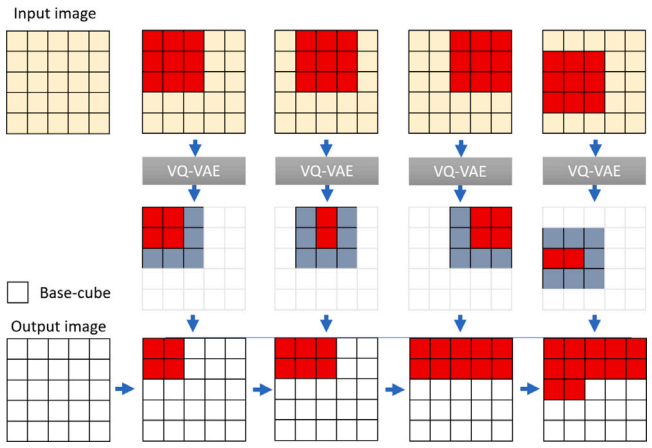



**Fig. 3. 2D Merging strategy with base-cube** – The use of convolutional padding can lead to border artifacts, see Fig. 2(a). To avoid this problem, we discard all the cubes that border unseen cubes (the discarded cubes is shown in gray color). This allows the resulting merged image to be continuous as if we process the entire large image at once without splitting.

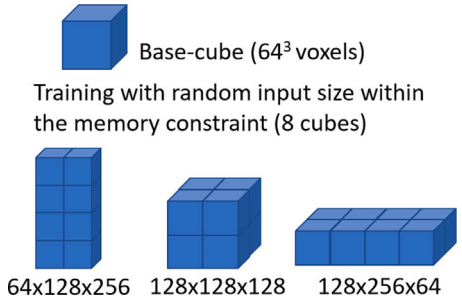



**Fig. 4. Training with multiple base-cubes**.

image reconstruction and processing procedures in order to avoid systematic problems. Due to resolution limitations of current imaging technology (field of view versus spatial resolution), porosity images from X-ray micro-CT images often contains a non-neglectful amount of sub-resolution porosity, i.e., the pores in these regions are below the image resolution. Therefore using porosity images provides an objective way of normalizing our training data between 0 (pore) – 100 (solid) and preserving the unresolved micro-porosity (1–99). The micro-porosity is critical when working with complex rocks, e.g., carbonates, where a big part of the image is represented by under-resolution pore structures.

For each training step, we make a random crop from a larger 3D image with a size variation between 50% to 200% of the original input size. Then, we resize the image with a tri-linear interpolation for up-sampling and porosity averaging for down-sampling. The use of random cropping as data augmentation reflects the variation in resolution and grain sizes when working with rock images.

*Validation metrics*

In order to validate the quality of the generated images, we compare the following petrophysical and topological properties to those of real rock images.

- **Porosity** ($\phi$): is the ratio between the pore and the total volume of the porous material, i.e., capacity to hold a fluid.
- **Permeability** ($k$): measures the hydraulic conductance of the media, i.e., under the same driving forces, a higher permeability means more fluid flow. We calculated the permeability by performing a fluid simulation on a pore-network model extracted from the 3D image using the method described in Ruspini et al. (2017), Øren et al. (2019), Ruspini et al. (2021).
- **Euler characteristic** ($\chi$): is a dimensionless functional describing the topology and connectivity of a structure, i.e., the relation between the number of disconnected components and the number of in-equivalent loops (Vogel, 2002). We calculate this number using the algorithm proposed in Blasquez and Poiraudeau (2003) and considering a 26-neighborhood connection.

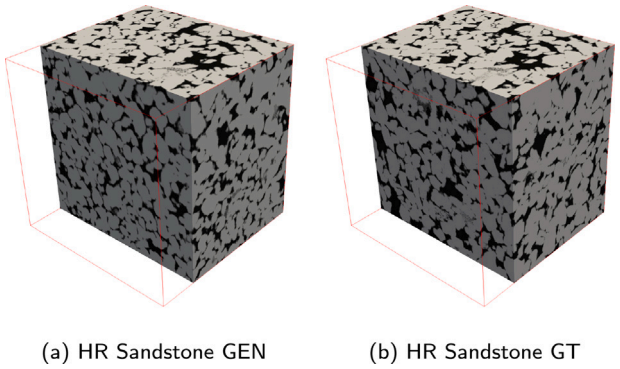



**Fig. 5. High-resolution (HR) Sandstone (1024 × 1024 × 1024 voxels)** – Generated image from a single high-resolution Sandstone image (top face) and the corresponding ground-truth.

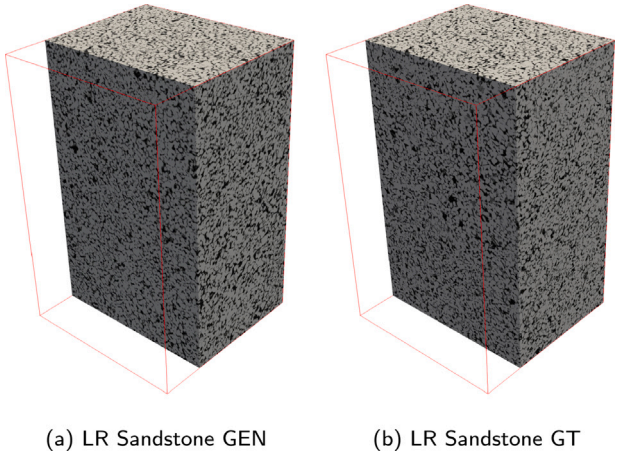



**Fig. 6. Low-resolution (LR) Sandstone (1250 × 1250 × 2000 voxels)** – Generated image from a single low-resolution Sandstone image (top face) and the corresponding ground-truth.

**Table 1**
Evaluation data – Our evaluation dataset comprises 4 images scanned at different resolutions. Resolution (Res.) in low resolution (LR) and high resolution (HR) is scanning resolution measured in μm per a voxels length.

| Rock Type | Voxel size | | | Res. |
|---|---|---|---|---|
| | X | Y | Z | [μm] |
| 1. HR Sandstone (Bentheimer) | 1024 | 1024 | 1024 | 1.94 |
| 2. LR Sandstone (Berea) | 1250 | 1250 | 2000 | 4.74 |
| 3. HR Carbonate (Reservoirs) | 1200 | 1200 | 1500 | 3.33 |
| 4. LR Carbonate (Reservoirs) | 1500 | 1500 | 3000 | 15.06 |

## 4. Results and discussion

To evaluate our workflow, we use 4 validation rock samples with different properties, resolutions, and image sizes, as shown in Table 1. We have generated a 3D image from the top 2D slice of each sample. The generated images and their corresponding GT are visualized with

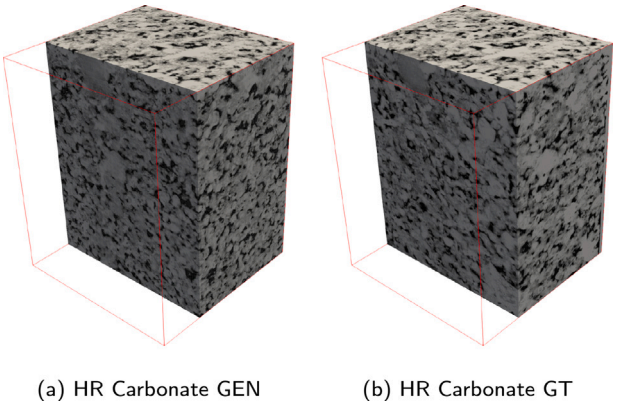



**Fig. 7. High-resolution (HR) Carbonate (1200 × 1200 × 1500 voxels)** – Generated image from a single high-resolution carbonate image (top face) and the corresponding ground-truth.

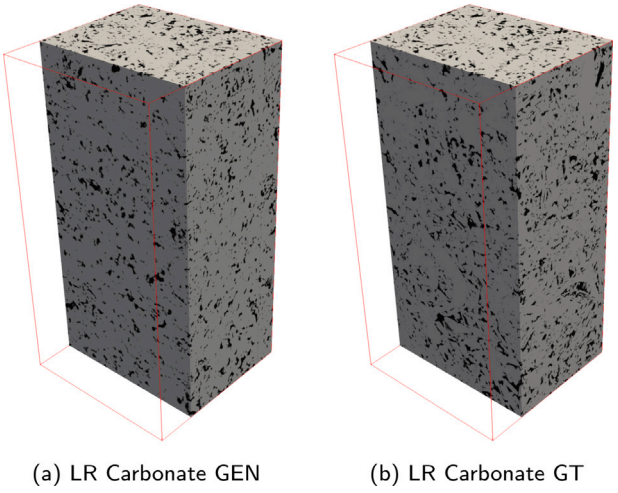



**Fig. 8. Low-Resolution (LR) Carbonate (1500 × 1500 × 3000 voxels)** – Generated image from a single low-resolution carbonate image (top face) and the corresponding to 3D ground-truth.

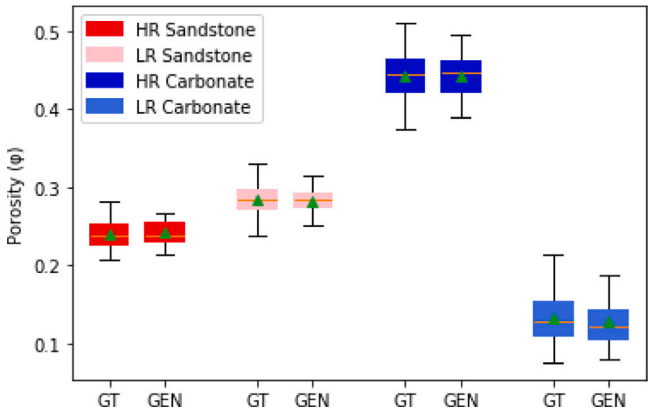



**Fig. 9. Porosity comparison –** The generated images (GEN) shows similar porosity distributions than the ground truths (GT) in all cases. We measure the porosity variation by dividing our image into several 200 × 200 × 200 voxels cubes and calculate the porosity for each of them. The results are plotted as a boxplot.

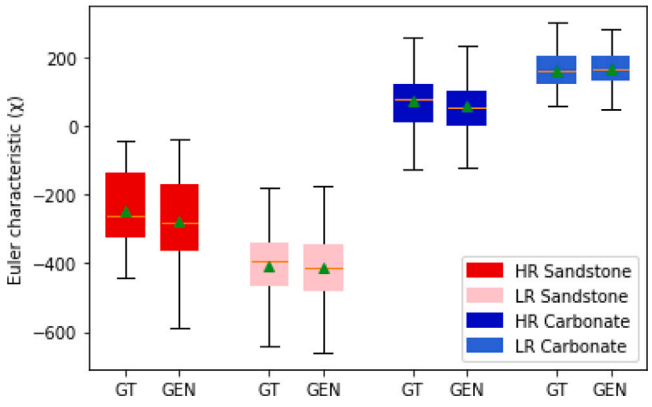



**Fig. 10. Euler number comparison –** We measure the Euler number distribution on several 200 × 200x200 voxels binary thresholded cubes taken from the larger image. The generated images (GEN) shows similar Euler number distribution to the ground truths (GT) in all cases.

Paraview (Ahrens et al., 2005) and shown in Figs. 5–8. As shown in these figures, our method is able to generate a completely different image with the same 2D input for each sample, while visually preserving the overall structure (e.g. grain sizes/shapes, porosity). In Figs. 9 and 10, we compare porosity and Euler characteristic for the generated and ground-truth images. In order to analyze the variation of these properties, we have divided the image into multiple $200^3$ voxels cubes and plot the distributions of their properties. In addition, we have calculated the permeability in all three directions for each of the sub-samples to build porosity vs. permeability plots. The slope and dispersion in these plots reflect different types of porous media (Ruspini et al., 2017, 2021). Despite the fact that permeability is quite sensitive to the local variations in both volume and shape of the pores, the generated images and real images yield similar trends for all the different types of rocks.

Given that our generation method is based on autoregressive sampling, each new generation produces a unique image. To demonstrate this, in Fig. 12, we show several realizations that are generated from the same input image. This is an important advantage of the new method since it allows estimation of the properties variations within a given rock type using just a single 2D image.

A common problem when working with transformers, and autoregressive models in general, is output degeneration (Holtzman et al., 2019). A large-size generation often yields repetitive or unnatural results. In order to analyze the degeneration effect, we have generated an 8192 voxels deep image from a 512 × 512 2D input, shown in Fig. 13. Then, we calculated porosity and Euler characteristic along the generated direction using a rolling window of 512 × 512 × 256 voxels in size and 64 voxels in step-length. By looking at the image and the properties, no sign of degradation is observed even when the generated image is 32 times longer than the model input length.

One of the main motivations for using VQ-VAE as a framework to the transformer is to significantly reduce computational cost. In Table 2, we show that the transformer step is the most time-demanding part of the entire workflow. It is important to consider that even though the input image size is a $1024^3$ voxels cube, the size that we need to sample using our transformer is only $64^3$ voxels, thanks to the encoder-quantization step. We estimate that generating a full-size $1024^3$ cube image using only a transformer would take around 4 years with the same hardware. Moreover, using just a transformer in a generation would also require a bigger receptive field, i.e, longer input sequence length, to capture the image context.

When it comes to determining the input sequence length of the model, the time constraint is an important factor to consider, as shown in Fig. 14. Another limiting factor is the memory and computation required to train a transformer, which grows quadratically with the sequence length (Beltagy et al., 2020). Since we are working with 3D images, doubling the input image size in each direction increases the input sequence length by 8-fold, thus massively increasing the generation time and memory consumption. On the other hand, having

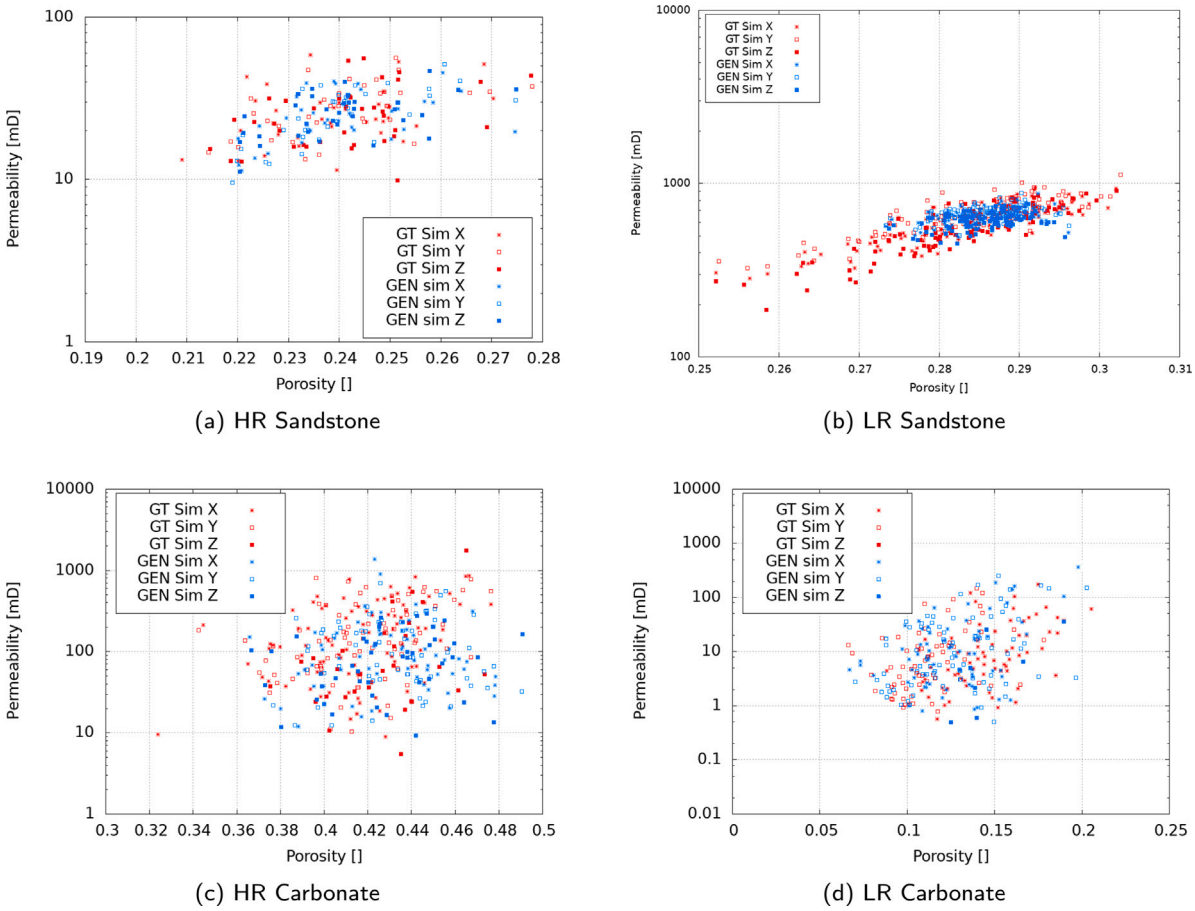



**Fig. 11. Permeability vs. porosity trends** – We calculate permeability in all directions (X,Y,Z) using pore-network simulations on the sub-samples of each rock type. The results show similar trends in permeability versus porosity between the generated images and the corresponding GTs. In Fig. 11b, the GT samples have a wider range of porosity and permeability than the GEN sample, this is due to the GEN sample only using information from the 2D input to construct the 3D image thus preventing large variation in image properties along the z-axes as shown in Fig. 13. However, this is not the case for some real rock samples where the properties of different image parts might vary significantly.

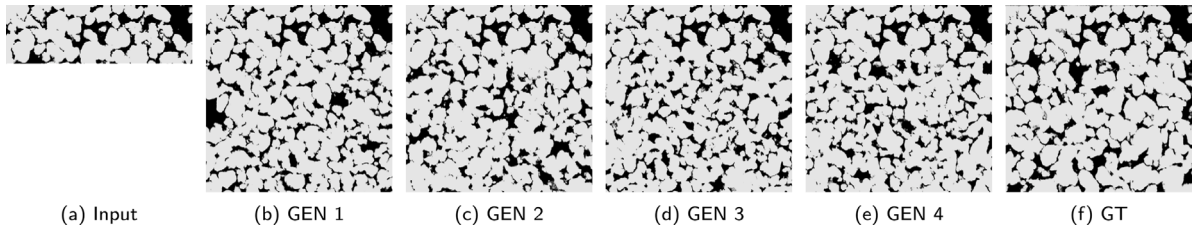



**Fig. 12. 2D vertical slices along the central axis of different 3D realizations and the GT** – An advantage of transformer-based generation methods is the ability to sample different realizations from a given input. This figure show that given 1/4 of the ground truth image as conditions, each new generation has produced a unique completion of the image. However all the four generated images still have a similar rock structure and properties, such as grain size and pore distribution. This ability could enable the user to predict the distribution in physical properties of a certain type of rock with just a single input.

a longer input sequence length should, in most cases, improve the generation quality (Parmar et al., 2018). Therefore the optimal solution will always be delimited by these mentioned factors. In our case, we have chosen to work with an input length of 16 × 16×10 due to the memory limitation of our GPU (Graphics Processing Unit).

## 5. Conclusion

In this work, we have proposed a workflow that combines VQ-VAE, Image Transformer, and size-invariant GAN, to solve the problem of large 3D image generation from a single 2D rock image. To validate our workflow results, we have generated several large-size 3D images of different resolutions from 4 different types of rocks. We then compared petrophysical and topological properties to those obtained for a real rock image and found encouraging results. Since acquiring 2D images is significantly cheaper and faster than 3D image acquisition, this workflow could potentially reduce the cost of rock's physical properties

**Table 2**
Time spent for each step – The wall time taken to generate a $1024^3$ voxel cube with a RTX3090. The sampling process with the transformer (using a 16 × 16 × 10 input sequence length) is the most time-consuming step. A shorter generation time could potentially be achieved with proper optimization and model pruning.

| Step | Wall Time |
|---|---|
| Encode and quantization | 38 s |
| Sample with transformer | 8 h 54 min 53 s |
| De-quantization and decode | 1 min 33 s |
| GAN | 1 min 1 s |

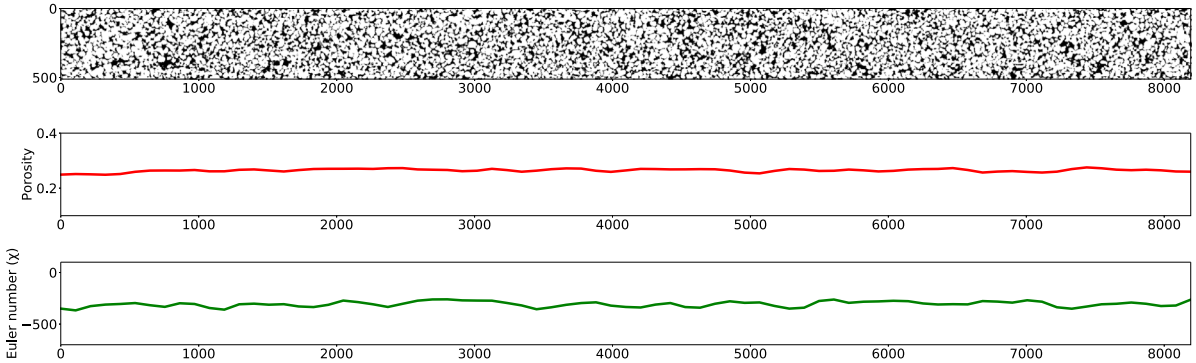



**Fig. 13. Variation in properties of an 8192 voxels deep generated 3D image from a 512 × 512 2D input – Top image:** The rotated 2D slice along the central axis of the generated image. **Middle plot:** Porosity measured using rolling windows of 512 × 512 × 256 and 64 in step-length along the depth direction. **Bottom plot:** Euler number measured using rolling windows of 512 × 512 × 256 in size and 64 in step-length along the depth direction.

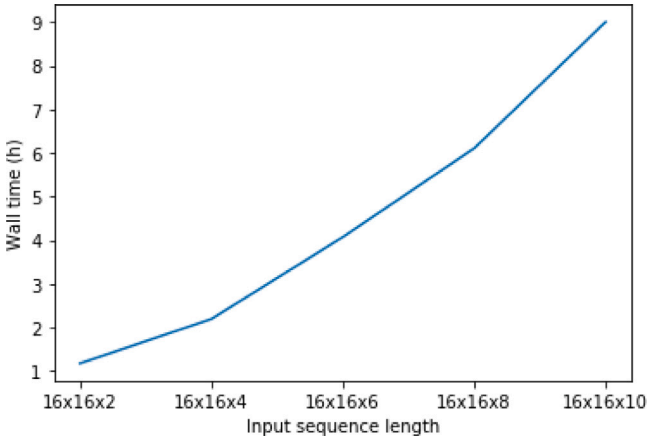



**Fig. 14. Input sequence length of the transformer is by far the most impacting factor when it comes to generation time** - The graph shows relationship between input sequence length and time needed to generate a sequence of 262 144 length ($64^3$) encoded and quantized from an $1024^3$ voxels image cube.

characterization which is crucial in many industries, such as CO2 capture and storage, hydrology, oil & gas.

*Limitations*

Since we only train with porosity maps in this work, we would need to add an extra step of converting other data-type into porosity maps in order for the model to work. However, this conversion can, in most cases, be done with image thresholding or existing segmentation tools. When dealing with noisy/artifacts in input images, we advise using a separate model for noise filtering as pre-process even though it is possible to incorporate noise filtering capacity into our model. To incorporate noise filtering capacity into our model, we would need to make the three parts of our model noise-robust, since they all take the 2D image as input. This could also have a negative impact on the performance of the generative model since it would have to learn another task.

This method is specifically built to solve the 2D to 3D generation problem of rock images and therefore is expected to work best in this domain. Based on our understanding, we expect the method to perform well on problems that are strictly defined by some physical or statistical distribution, such as material science or biology. However, we have not tested our method on other domains since it is outside the scope of this work.

## CRediT authorship contribution statement

**Johan Phan:** Designed the research, Developed the network architecture and performed the experiments, Wrote the manuscript, Reviewed the manuscript. **Leonardo Ruspini:** Designed the research, Wrote the manuscript, Reviewed the manuscript. **Gabriel Kiss:** Reviewed the manuscript. **Frank Lindseth:** Reviewed the manuscript.

## Declaration of competing interest

One or more of the authors of this paper have disclosed potential or pertinent conflicts of interest, which may include receipt of payment, either direct or indirect, institutional support, or association with an entity in the biomedical field which may be perceived to have potential conflict of interest with this work. For full disclosure statements refer to https://doi.org/10.1016/j.petrol.2022.110648. Johan Phan reports financial support was provided by Petricore Norway. Leonardo Ruspini reports financial support was provided by Eni SpA. Leonardo Ruspini reports financial support was provided by Respol. Leonardo Ruspini reports financial support was provided by Chevron Corp. Leonardo Ruspini reports a relationship with Eni SpA that includes: consulting or advisory, funding grants, and non-financial support. Leonardo Ruspini reports a relationship with Australian National Low Emissions Coal Research and Development that includes: consulting or advisory.

## Acknowledgments

This work was partially supported by the Norwegian Research Council (grant number 296093) and the members of the SmartRocks joint industry project (ENI SpA, Repsol AS, and Chevron Corporation).

## Appendix A. Model architectures

The architectures of the VQ-VAE and the GAN models use in this work are shown in Figs. 15 and 16. For the transformer model, we use the GPT-3 architecture without any modification.

## Appendix B. Effect of transformer receptive field

In the Results and Discussion section, we show that increases in the size of the receptive field are at least linearly proportional to the generation/inference time of the transformer. However, the size of the receptive field also has a significant effect on the generation quality of our model, as shown in Figs. 17 and 18. The results in these figures indicate that a smaller receptive field increases the likelihood of generation degradation.

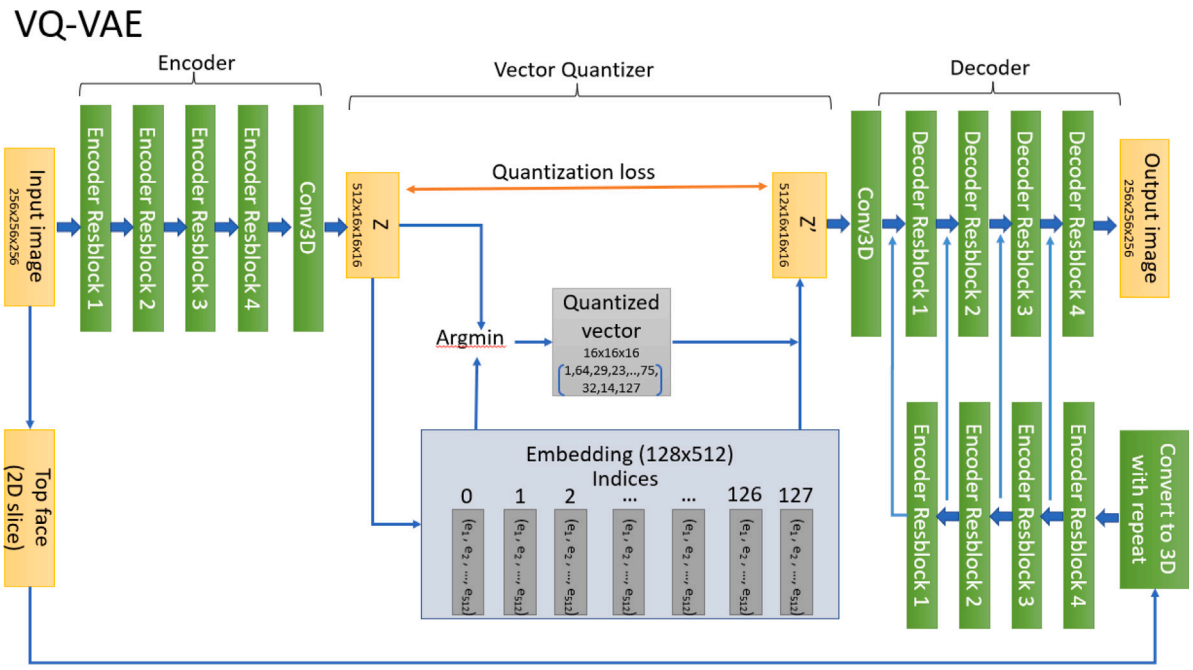



**Fig. 15. VQ-VAE architecture** – The VQ-VAE model used in this work comprises an encoder and a decoder with 4 residual blocks each. Each residual block has 4 convolution layers with PReLU activation. To discretize the encoded output (Z vector), we use an embedding dictionary with 128 indices and an embedding dimension of 512. We replace each value in the encoded vector with the dictionary index that has the closest embedding value to create the Z' vector, where Z' is equal to Z in case of lossless quantization. In addition to MSE (mean square error) loss between the input and the output images, quantization loss, defined as the absolute distance between the Z and Z', is also used. Finally, the first 2D slice (top face) of the input image is also used as an additional input for every decoder block.

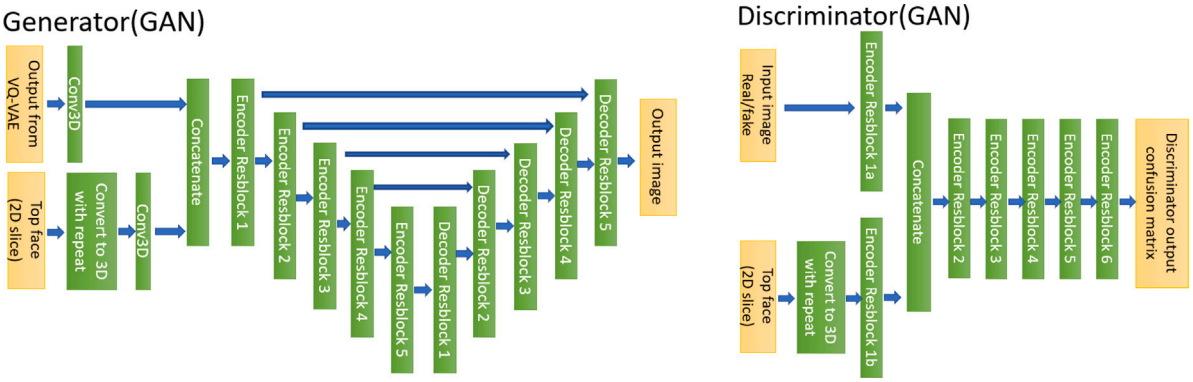



**Fig. 16. GAN architecture** – Our GAN models take the output of the VQ-VAE and the 2D image as input. To concatenate the 3D image with the 2D image, we repeat the 2D image into a 3D stack with the same size as the 3D input. Similar to the VQ-VAE, each residual block of the GAN models has 3 convolutional layers with PReLU activation. The residual decoder block, in our case, uses pixel-shuffle for upsampling instead of de-convolutional to reduce the checkerboard artifact of GAN.

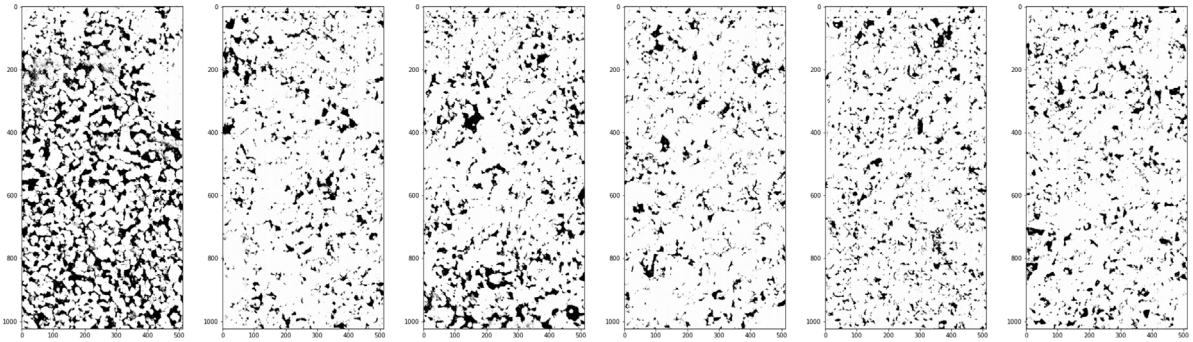



**Fig. 17. The effect of transformer receptive field on images generated along the Z direction (the figures show XZ plane slice)** – a small receptive field (16 × 16×2 to 16 × 16 × 6 voxels) increases the likelihood of output degeneration. However, the improvement can be minimal when the receptive field is above a certain size, in this case, larger than 16 × 16 × 8 voxels.

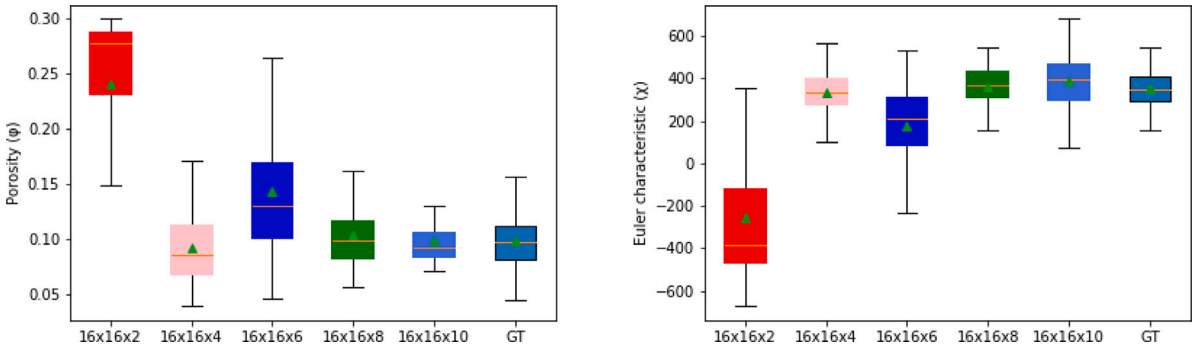



**Fig. 18. Properties comparison between results generated with different transformer receptive field** – As shown in Fig. 12, the images generated with a receptive field of 16 × 16 × 2 voxels and 16 × 16 × 6 voxels have suffered from degradation along the $Z$-axis, thus resulting in a significant difference in the calculated porosity and Euler characteristic in these cases compared to the GT and other cases.

## References


Adler, P.M., Jacquin, C.G., Quiblier, J.A., 1990. Flow in simulated porous media. Int. J. Multiph. Flow. 16 (4), 691–712.

Ahrens, James, Geveci, Berk, Law, Charles, 2005. Paraview: An end-user tool for large data visualization. Vis. Handb. 717 (8).

Arns, C., Sakellariou, A., Senden, T., Senden, T., Sheppard, A., Knackstedt, M., 2003. Petrophysical properties derived from X-ray CT images. APPEA J. 43, 577–586.

Bakke, S., Øren, P.E., 1997. 3D pore-scale modeling of sandstones and flow simulations in the pore networks. SPE J. 2, 136–149.

Beltagy, Iz, Peters, Matthew E., Cohan, Arman, 2020. Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.

Blair, Stephen C., Berge, Patricia A., Berryman, James G., 1996. Using two-point correlation functions to characterize microgeometry and estimate permeabilities of sandstones and porous glass. J. Geophys. Res.: Solid Earth 101 (B9), 20359–20375.

Blasquez, Isabelle, Poiraudeau, Jean-François, 2003. Efficient processing of Minkowski functionals on a 3D binary image using binary decision diagrams. In: WSCG. UNION Agency-Science Press.

Brown, Tom B, Mann, Benjamin, Ryder, Nick, Subbiah, Melanie, Kaplan, Jared, Dhariwal, Prafulla, Neelakantan, Arvind, Shyam, Pranav, Sastry, Girish, Askell, Amanda, et al., 2020. Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Bruns, S., Stipp, S.L.S., Sørensen, H.O., 2017. Statistical representative elementary volumes of porous media determined using greyscale analysis of 3D tomograms. AWR.

Čapek, P, Hejtmánek, V, Brabec, L, Zikánová, A, Kočiřík, M, 2009. Stochastic reconstruction of particulate media using simulated annealing: improving pore connectivity. Transp. Porous Media 76 (2), 179–198.

Chen, Mark, Radford, Alec, Child, Rewon, Wu, Jeffrey, Jun, Heewoo, Luan, David, Sutskever, Ilya, 2020. Generative pretraining from pixels. In: International Conference on Machine Learning. PMLR, pp. 1691–1703.

Coiffier, Guillaume, Renard, Philippe, Lefebvre, Sylvain, 2020. 3D geological image synthesis from 2D examples using generative adversarial networks. Front. Water 2, 30.

Fan, Angela, Lewis, Mike, Dauphin, Yann, 2018. Hierarchical neural story generation. arXiv preprint arXiv:1805.04833.

Feng, J., He, X., Teng, Q., Ren, C., Chen, H., Li, Y., 2019. Accurate and fast reconstruction of porous media from extremely limited information using conditional generative adversarial network. arXiv:1905.02135.

Feng, Junxi, Teng, Qizhi, Li, Bing, He, Xiaohai, Chen, Honggang, Li, Yang, 2020. An end-to-end three-dimensional reconstruction framework of porous media from a single two-dimensional image based on deep learning. Comput. Methods Appl. Mech. Engrg. 113043.

Golab, A., Knackstedt, M., Averdunk, H., Senden, T., Butcher, A., Jaime, P., 2010. 3D porosity and mineralogy characterization in tight gas sandstones. Lead. Edge 936–942.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, Bengio, Yoshua, 2014. Generative adversarial nets. Adv. Neural Inf. Process. Syst. 27.

Gupta, Ankit, Dar, Guy, Goodman, Shaya, Ciprut, David, Berant, Jonathan, 2021. Memory-efficient transformers via top-k attention. CoRR, abs/2106.06899.

Holtzman, Ari, Buys, Jan, Du, Li, Forbes, Maxwell, Choi, Yejin, 2019. The curious case of neural text degeneration. arXiv preprint arXiv:1904.09751.

Islam, Md. Amirul, Kowal, Matthew, Jia, Sen, Derpanis, Konstantinos G., Bruce, Neil D.B., 2021. Position, padding and predictions: A deeper look at position information in CNNs. CoRR, abs/2101.12322.

Karras, Tero, Laine, Samuli, Aittala, Miika, Hellsten, Janne, Lehtinen, Jaakko, Aila, Timo, 2020. Analyzing and improving the image quality of stylegan. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119.

Kayhan, Osman Semih, van Gemert, Jan C., 2020. On translation invariance in CNNs: Convolutional layers can exploit absolute spatial location. CoRR, abs/2003.07064.

Mosser, Lukas, Dubrule, Olivier, Blunt, Martin J., 2017. Reconstruction of three-dimensional porous media using generative adversarial neural networks. Phys. Rev. E 96 (4).

Mosser, Lukas, Dubrule, Olivier, Blunt, Martin J., 2018. Stochastic reconstruction of an oolitic limestone by generative adversarial networks. Transp. Porous Media 125 (1), 81–103.

Okabe, Hiroshi, Blunt, Martin Julian, 2004. Prediction of permeability for porous media reconstructed using multiple-point statistics.. Phys. Rev. E 066135, 70 6 Pt 2.

Oord, Aaron van den, Kalchbrenner, Nal, Vinyals, Oriol, Espeholt, Lasse, Graves, Alex, Kavukcuoglu, Koray, 2016. Conditional image generation with pixelcnn decoders. arXiv preprint arXiv:1606.05328.

Oord, Aaron van den, Vinyals, Oriol, Kavukcuoglu, Koray, 2017. Neural discrete representation learning. arXiv preprint arXiv:1711.00937.

Øren, P.E., Bakke, S., 2002. Process based reconstruction of sandstones and predicition of transport properties. Transp. Porous Media 46, 311–343.

Øren, P.E., Ruspini, L.C., Saadatfar, M., Sok, R.M., Knackstedt, M., Herring, A., 2019. In-situ pore-scale imaging and image-based modelling of capillary trapping for geological storage of CO2. Int. J. Greehouse Gas Control.

Pant, Lalit M., 2016. Stochastic characterization and reconstruction of porous media. Phys. Rev. E.

Parmar, Niki, Vaswani, Ashish, Uszkoreit, Jakob, Kaiser, Lukasz, Shazeer, Noam, Ku, Alexander, Tran, Dustin, 2018. Image transformer. Int. Conf. Mach. Learn. 4055–4064.

Razavi, Ali, van den Oord, Aaron, Vinyals, Oriol, 2019. Generating diverse high-fidelity images with vq-vae-2. Adv. Neural Inf. Process. Syst. 14866–14876.

Ruspini, L.C., Bakke, S., Øren, P.E., 2021. Multiscale digital rock analysis for complex rocks. Transp. Porous Media 139, 301–325.

Ruspini, L.C., Farokhpoor, R., Øren, P.E., 2017. Pore-scale modeling of capillary trapping in water-wet porous media: A new cooperative pore-body filling model. Adv. Water Resour. 108, 1–14.

Strebelle, S., 2002. Conditional simulation of complex geological structures using multiple-point statistics. Math. Geol. 34, 1–21.

Tahmasebi, P., Hezarkhani, A., Sahimi, M., 2012. Multiple-point geostatistical modeling based on the cross-correlation functions. Comput. Geosci. 16, 779–797.

Tahmasebi, P., Sahimi, M., Caers, J., 2014. MS-CCSIM: accelerating pattern-based geostatistical simulation of categorical variables using a multi-scale search in Fourier space. Comput. Geosci. 67, 75–88.

Thanh-Tung, Hoang, Tran, Truyen, 2020. Catastrophic forgetting and mode collapse in GANs. In: 2020 International Joint Conference on Neural Networks, IJCNN. IEEE, pp. 1–10.

Valsecchi, Andrea, Damas, Sergio, Tubilleja, Cristina, Arechalde, Javier, 2020. Stochastic reconstruction of 3D porous media from 2D images using generative adversarial networks. Neurocomputing 399, 227–236.

Vogel, Hans-Jörg, 2002. Topological characterization of porous media. In: Morphology of Condensed Matter. Springer, pp. 75–92.

Volkhonskiy, D., Muravleva1, E., Sudakov, O., Orlov, D., Belozerov, B., Burnaev, E., Koroteev, D., 2019. Reconstruction of 3D porous media from 2D slices. arXiv:1901.10233v4.

Yan, Wilson, Zhang, Yunzhi, Abbeel, Pieter, Srinivas, Aravind, 2021. VideoGPT: Video generation using VQ-VAE and transformers. arXiv preprint arXiv:2104.10157.

Zhang, Fan, Teng, Qizhi, Chen, Honggang, He, Xiaohai, Dong, Xiucheng, 2021. Slice-to-voxel stochastic reconstructions on porous media with hybrid deep generative model. Comput. Mater. Sci. 186, 110018.

Zhao, Jiuyu, Wang, Fuyong, Cai, Jianchao, 2021. 3D tight sandstone digital rock reconstruction with deep learning. J. Pet. Sci. Eng. 207, 109020.