Rock Image Super-resolution: From CT to micro-CT

Zitong Huang and Minghui Xu
Department of Energy Science and Engineering
Stanford University
zthuang@stanford.edu, minghuix@stanford.edu

Abstract

Understanding the subsurface plays a pivotal role in the sustainable energy transition through large-scale initiatives such as underground carbon sequestration and hydrogen storage. Visualizing rock structures and pore networks is important for understanding fluid–rock interactions and serves as a critical first step in implementing these geologic projects. X-ray Computed Tomography (CT) is well-known for providing rock imaging at approximately 100 μm with high temporal resolution. This makes it suitable for in situ observations of fluid–rock interactions during experiments. However, the limited spatial resolution of CT introduces significant uncertainty in key pore-scale features that govern transport processes, such as fracture locations, pore connectivity, and micro-scale pores or fractures. As a robust alternative, micro-CT offers around tenfold improvement in spatial resolution but is constrained by high acquisition costs in both time and resources. This limits its use for in situ dynamic studies. To address this trade-off between spatial and temporal resolution, this study proposes a workflow that leverages super-resolution X-ray CT, powered by various computer vision models, to efficiently generate high-resolution micro-CT rock images from fast-acquired CT images. The outcomes of this work have the potential to significantly improve the economic viability and implementation feasibility of large-scale subsurface energy projects.

1. Introduction

Understanding the pore structure of rocks is important for large-scale geological and reservoir engineering applications, including CO_2 sequestration, underground hydrogen storage, and geothermal energy recovery. To understand the subsurface characteristics, reservoir engineers need to take rock samples from the subsurface reservoir of interest and perform lab experiments on the collected samples. To obtain the properties of the rock and monitor the experiment progress, 3-D CT scans on the sample before, during, and after the experiments are essential. For a 5 cm-long rock sample, a 3-D CT scan can be obtained in around 3 minutes, but a 3-D micro-CT scan requires at least 3 hours [8]. This long image acquisition time limits the micro-CT's ability to perform in situ observation (i.e., observation during the experiment). On the other hand, the resolution of micro-CT images (around 10 μm) is significantly finer than CT images (around 100 μm), and it can give a detailed characterization of the pore networks of rocks. To take advantage of both short data-acquisition time and high image resolution, we plan to translate X-ray CT images of rock to micro-CT in this project.

The fundamental idea is to treat this task as micro-CT generation conditioned on paired CT scans and spatial masks that discriminate between regions inside and outside the rock. Previous research on using generative adversarial networks (GANs) for this conditional generation has been explored in Murugesu et al. [8]. However, recent advances in deep generative models show that diffusion models and flow matching models achieve state-of-the-art performance in many computer vision tasks [2] and are much easier to train compared to GANs. Denoising diffusion probabilistic models [3, DDPMs] are one of the most famous diffusion models that learn and reproduce complicated distributions of high-dimensional data by maximizing the likelihood of training data. The backbone of diffusion models, U-Net, was later replaced by transformers called diffusion transformer [9, DiT] and scalable interpolant transformer [7, SiT], which make the model more scalable. Therefore, we will primarily focus on two popular models—DDPMs and SiT—and compare them in terms of generation efficiency and performance.

2. Related Work

Several previous studies have investigated the super resolution of computed tomography images using deep-learning based approaches. However, instead of transferring among modalities (i.e., from CT to micro-CT), previous studies have largely focused on recovering high-resolution structure within a single imaging modality with moderate scaling factors and either paired or reference-based supervision.

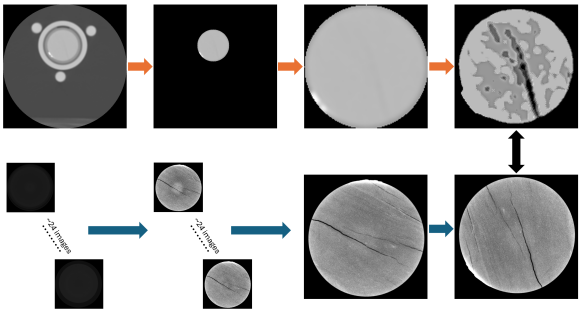

Figure 1. Visualization of the processing stages for a pair of CT (top row: rock matrix identification, cropping, and normalization) and micro-CT (bottom row: normalization, averaging, and alignment) images from the dataset described in Section 4.

For instance, You et al. [14] introduced GAN-CIRCLE to enhance tibia CT using an ensemble of adversarial and cycle-consistency losses. They successfully enhanced the resolution from 300 μm to 150 μm. To better capturing long-range dependencies and global contextual relationships, Zhou et al. [16] implemented a joint denoising scheme that combines GAN with reference-based transformer approaches to enhance CT scans from low to high resolution. In the context of geology, Liu et al. [5] addressed unpaired rock CT super resolution using a non-local CycleGAN. They successfully reached 4 times resolution enhancement for CT images. Meanwhile, Ma et al. [15] applied diffusion models to synthetically enhance micro-CT carbonate volumes by up to 16 times. Despite the large resolution enhancement, their model assumed already micro-scale input. These studies operate within the same modality and improve from already decent-resolution inputs. In contrast, the task of predicting micro-CT-scale features (< 10 μm) directly from CT (> 100 μm) introduces a more severe resolution gap, compounded by cross-modal discrepancies and the lack of voxel-aligned training pairs.

The only known attempt to enhance the resolution of rock images across modality is done by Murugesu et al. [8]. They employed conditional GAN to enhance the resolution of rock CT iamges. However, their results revealed two main limitations. First, due to memory constraints and the limitations of CNN-based architectures, both CT and micro-CT images were resized to 256 $\times$ 256, effectively downscaling the micro-CT by a factor of four and upscaling the CT by a factor of two. While they still preserved a relative resolution gap, it compromised physical fidelity. Second, the super-resolution performance remains limited, with the best test sample achieving only an SSIM of 0.1999 and a PSNR of 18.286. This indicates that the generated micro-CT images still struggle to recover fine-scale structural details essential for downstream geological analysis. This project uses the same raw dataset as Murugesu et al. [8], but employs more advanced computer vision models, namely DDPMs and SiT.

Most previous studies utilized GAN-based models. Recent work has shown that diffusion models offer superior training stability and mode coverage compared to GANs by gradually learning to reverse a fixed Markovian noise process through likelihood maximization [3]. This avoids the challenges of adversarial minimax optimization while producing more diverse samples with stronger theoretical guarantees for convergence.

Building on these advantages, several diffusion model formulations have been proposed to further improve image synthesis performance. Among them, DDPMs [3], Score-Matching Langevin Dynamics (SMLD) [12], and score-based Stochastic Differential Equations (SDEs) [13] represent three complementary perspectives on diffusion modeling. Song et al. [12] demonstrated that forward and reverse SDEs can describe the noise injection and denoising processes, and derived the corresponding SDE formulations for DDPMs and SMLD, referred to as the Variance Preserving (VP) and Variance Exploding (VE) SDEs, respectively. In parallel, flow-based generative models [4, 6] and recent work on unifying diffusion and flow-matching frameworks through stochastic interpolants [1] have explored alternative approaches to generative modeling. These methods provide a theoretical bridge between score-based diffusion and deterministic flow models, and have shown strong potential for high-quality image super-resolution.

3. Problem Statement

Specifically, the input for the model contains three channels. The first two are conditioning CT and the same-sized Gaussian noise images. As shown in Fig. 1, all rock cores in the CT and micro-CT images have a circular shape, so we introduced a third masking channel of pixel values of 0 outside and 1 inside the circle to guide the network to focus only on the region that contains the rock matrix. The output is the corresponding micro-CT. In this study, the CT-micro-CT image pairs we used were obtained from shale rock. Model performance is evaluated qualitatively, through visual comparison of generated and ground-truth micro-CT, and quantitatively, using Peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and customized loss. We analyzed stability trends across timesteps and epochs. This project is connected to our broader research in geologic imaging, but the model architecture design, training strategy, and evaluation framework were uniquely developed.

4. Dataset

We have obtained the 3-D CT and micro-CT scans of shale rock samples. Using a series of customized programs, we have developed a workflow to extract, process, and cor-

relate the CT images with corresponding micro-CT images. In brief, we first read the raw CT and micro-CT images from their directories and handled DICOM-to-PNG conversion for CT. The 3-D volumetric scans were converted to a stack of 2-D images, and the resolution of the CT and micro-CT images is 195 μm and 27 μm. For each CT and micro-CT pair, a tight circular mask was then created to isolate the circular rock sample from the background. To ensure consistency, pixel value intensities inside the circular region were normalized from 0 to 1; CT images underwent histogram matching to align their intensity distribution with micro-CT images. Each normalized core was then tightly cropped around the detected circle and resized (CT to 128×128 and micro-CT to 1024×1024). To further enhance the image quality and contrast of distinct features in the images, we implemented Gaussian noise filtering on the images. To account for the difference in spatial resolution, we averaged the pixel values of approximately 24 micro-CT slices to match the resolution of a single clinical CT slice. We also implemented Particle swarm optimization (PSO)-based alignment to correct rotational misalignment between CT and micro-CT images. A detailed visualization of the image processing workflow is illustrated in Fig. 1. One of the distinct features of shale rock is its long and thin fractures. Because fractures are pore spaces, their pixel values are usually close to 0. To attenuate the prediction of the distinct features, we inverted the normalized pixel values in the original processed image pairs, so pixels in the fractures will have the highest pixel values, and misprediction on these pixels will result in greater loss.

After a series of image processing and visual inspection, we selected 83 pairs of CT and micro-CT images. The data augmentation consists of rotating each image by 10 degrees, cropping the images evenly into 4, and vertically flipping the cropped images. So, we have 83×36×4×2=23,904 images, and we used 80% for training, 10% for validation, and 10% for the test set.

5. Method

5.1. Diffusion and flow matching

Diffusion models and flow matching are two popular and closely related generative methods, which include approaches such as DDPMs and SiT. We will first introduce the principles of DDPMs, followed by SiT, which unifies diffusion models and flow matching, as the latter has been shown to generate higher-fidelity images.

DDPMs include two essential processes: a pre-defined forward process and a trainable reverse process that generate data from Gaussian noise. The forward process converts the mirco-CT, which is sampled from training sets, into Gaussian noise by gradually adding small Gaussian noise step by step, as illustrated in Fig 2.

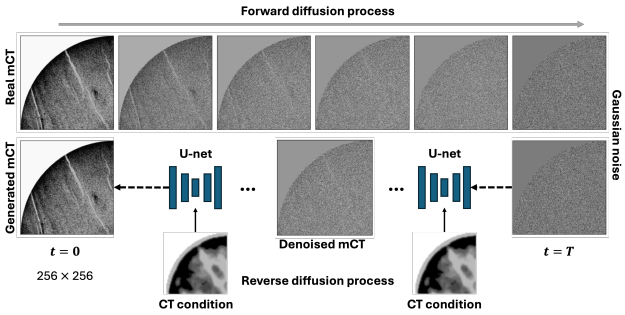

Figure 2. Illustration of the forward (noise perturbation) and reverse (denoising) processes in DDPM for micro-CT super-resolution conditioned on CT images.

The forward transformation follows a Gaussian distribution with mean $\sqrt{1-\beta_t}\,\mathbf{x}_{t-1}$ and variance $\beta_t\mathbf{I}$:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\,\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \quad t \in \{1, \ldots, T\},$$

where β_t represents the variance of the added noise level, and $\mathbf{I}$ represents the identity matrix, while t denotes the iteration number and T is the total number of iterations. With a sufficiently small value for β_t, the reverse process can be shown to have a function similar to the forward process [11, 3], resulting in the inverse conditional distribution $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ also being a Gaussian distribution. In Ho et al. [3], β_t is set to increase linearly, with $\beta_1 = 10^{-4}$ and $\beta_T = 0.02$.

To maximize the likelihood of the training data, the smoothed L1 loss function L_t is derived for training the neural network in DDPM, and is given by:

$$L_t(\boldsymbol{\epsilon}_\theta, \boldsymbol{\epsilon}_t) = \begin{cases} 0.5(\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{ct}, t) - \boldsymbol{\epsilon}_t)^2, & \text{if } |\boldsymbol{\epsilon}_\theta - \boldsymbol{\epsilon}_t| < 1 \\ |\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{ct}, t) - \boldsymbol{\epsilon}_t| - 0.5, & \text{otherwise} \end{cases}$$

where $\boldsymbol{\epsilon}_t$ is a standard Gaussian noise independently sampled at step t, and $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{ct}, t)$ represents the noise predicted by a U-Net parameterized by θ, conditioned on the CT image $\mathbf{ct}$.

Once the denoising U-Net is trained, standard Gaussian noise can be sampled and iteratively denoised using the network—conditioned on the low-resolution CT input—to generate a clean micro-CT image, as illustrated in Fig. 2.

SiT involves not only the diffusion or flow matching generation process but also leverages an autoencoder with Kullback–Leibler divergence (AutoencoderKL) for more efficient generation. The first step is to train the AutoencoderKL from scratch to convert the original micro-CT domain into latent features with lower dimensionality, while still allowing reconstruction of the original micro-CT images, as denoted by encoder and decoder in Fig 3. To enhance the reconstruction quality, we introduce two additional components: perceptual loss and a patch discriminator, following the procedure in [10]. Second, within the latent space, we build a flow-matching generative model

that uses a transformer to learn the velocity fields that drive Gaussian noise toward the latent micro-CT features, conditioned on the CT data.

The forward process of SiT is described by $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}$, where $\mathbf{x}_0$ is the original micro-CT image, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is standard Gaussian noise. The transformer is trained to predict the velocity field that reverses this process. The loss function for training the transformer with parameters θ is given by

$$L_{\mathbf{v}}(\theta) = \int \mathbb{E}\left[\|\mathbf{v}(\mathbf{x}_t, \mathbf{ct}, t) - \dot{\mathbf{x}}_t\|^2\right] \mathrm{d}t,$$

where $\mathbf{v}(\mathbf{x}_t, \mathbf{ct}, t)$ denotes the velocity predicted by the transformer, conditioned on the current state $\mathbf{x}_t$, the CT image $\mathbf{ct}$, and iteration step t.

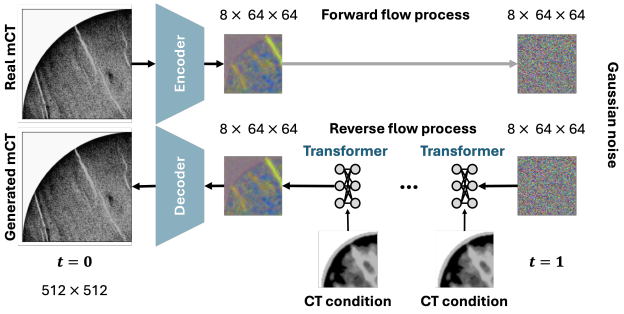

Figure 3. SiT-based super-resolution framework showing forward (noise perturbation) and reverse (generation) processes for producing micro-CT images conditioned on CT input.

After training the transformer-based velocity estimator, an ordinary differential equation (ODE) solver, specifically Dopri5, is used to generate the latent micro-CT features from standard Gaussian noise. These latent features are then passed through the decoder of the autoencoder to produce the final micro-CT image, as illustrated in Fig. 3.

5.2. Experimental setup

For the DDPMs training, on top of the smoothed L1 loss, the overall loss also consisted of a customized loss function termed fracture loss that emphasizes fine fractures and pore regions using spatial variance, brightness weighting, and gradient-based structural losses. Specifically, the loss generated a weighted smoothed L1 loss with the following equations that give greater weights on regions with low variance (e.g., homogeneous pixel distribution inside fractures) and high brightness (e.g., congregated pore pixels). In the weighting equations, the hyperparameter α controls the sensitivity to local variance, γ adjusts the emphasis on brighter structures, and δ modulates the overall sharpness and influence of the combined weights. The local variance is computed using a sliding average filter (i.e., $\mathrm{Var}(x) = \mathrm{AvgPool}(x^2) - (\mathrm{AvgPool}(x))^2$). The total weight is defined as the product of variance and brightness weights:

$$w_{\text{tot}} = \exp(-\alpha \cdot \mathrm{Var}(x)) \cdot (x + 0.1)^{\gamma}$$

These weights are combined with a sharpening exponent and normalized:

$$w_{i,c,h,w} = \frac{\left(w^{\text{tot}}_{i,c,h,w} \cdot \text{mask}_{i,c,h,w}\right)^{\delta}}{\sum_{h=1}^{H}\sum_{w=1}^{W}\left(w^{\text{tot}}_{i,c,h,w} \cdot \text{mask}_{i,c,h,w}\right)^{\delta} + \epsilon},$$

where i, c, h, and w denote the sample, channel, height, and width indices, respectively and ϵ is a small constant for numerical stability. Finally, the loss is computed as a weighted smooth L1 loss:

$$\mathcal{L}_{\text{fracture}} = \frac{1}{N}\sum_{i=1}^{N} \text{clamp}\left(w_i L_t^i, \text{max} = 5.0\right),$$

where N is the batch size.

For the training of both DDPMs and SiT, we used the AdamW optimizer with an initial learning rate of 5×10^{-5}. For DDPMs, a `ReduceLROnPlateau` scheduler was applied with a reduction factor of 0.5 and a patience of 20 epochs. Noise levels were scheduled using a linear beta scheduler over 800 diffusion timesteps. Regularization was applied through weight decay of 5×10^{-4}, and gradient norm clipping with a threshold of 0.7 was used to improve training stability. For the SiT, we used a `cosine decay` scheduler, and we did not implement any weight decay and gradient norm clipping, as this combination gave better results during both the training and testing.

6. Results

For training, we have built and trained a DDPM for this image super-resolution task. Fig. 4 shows the training and validation losses of the DDPMs case, in which no over-fitting is observed. The SiT model maintains similar training and validation loss throughout training, as illustrated in Fig. 5, indicating that overfitting does not occur as well. We quantitatively evaluated the performance of the network on 60 samples in the test dataset. Figs. 6 and 7 present the quantitative performance of the trained DDPMs on the test dataset. Figs. 8 illustrates the qualitative performance. The average PSNR and SSIM scores on these samples are 19.90 and 0.358, respectively. Despite not being extremely high, they are already better than the best values reported from the previous study on similar dataset (PSNR=18.29 and SSIM=0.199) [8]. There are four main reasons behind the intrinsic difficulty in predicting micro-CT with high quantitative metrics. First, PSNR and SSIM are highly sensitive to small pixel-level differences, which may not impact perceptual quality but still lower the scores. Second, these metrics do not always reflect human visual perception, especially in structurally complex images like micro-CT. Third, even after denoising or correction, sub-

tle artifacts in micro-CT predictions can significantly reduce PSNR and SSIM. Fourth, the CT image inherently lacks the high-frequency small-scale features contained in the micro-CT. Each of the four areas presents distinct challenges that limit metric-based performance gains through model design.

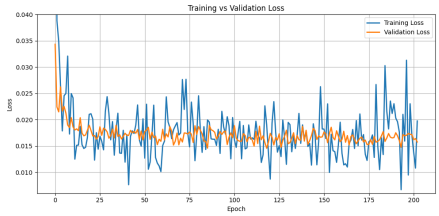

Figure 4. The training and validation loss for the DDPMs case. The loss is smoothed L1 loss + 1.5 $\times$ fracture loss.

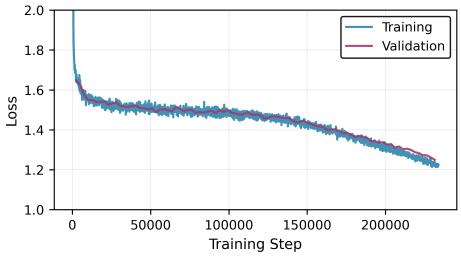

Figure 5. Training and validation loss curves of the SiT model. The loss represents the mean squared error of the velocity prediction.

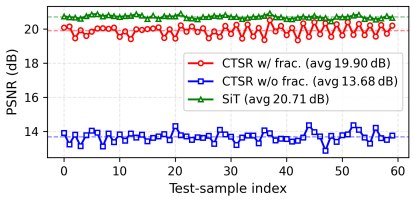

Figure 6. The peak signal-to-noise ratio for 60 images from the test dataset.

We treat the model trained without the fracture loss as the baseline, Figs. 6, 7, and 10 show both quantitative and qualitative improvement introduced by the loss.

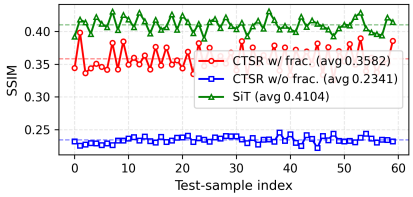

Figure 7. The structural similarity index measure for 60 images from the test dataset.

In addition to some examples with relatively high statistical indicators shown in Fig. 8, Fig. 9 displays five samples

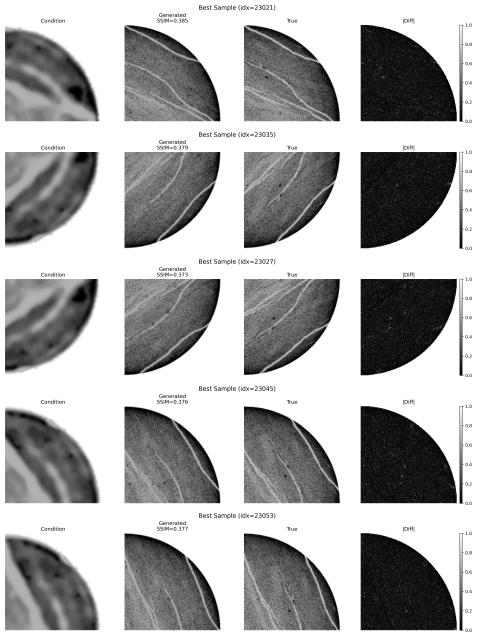

Figure 8. The 5 samples predicted by the DDPMs with fracture loss that achieved the highest SSIM among the 60 test images. Each row (from left to right) contains the conditioning CT, the generated micro-CT, the true micro-CT, and the absolute pixel-wise prediction error.

with the lowest SSIM scores. Although their SSIM scores are not as high as others, the key fractures are still generated at a fine resolution, matching well with the reference micro-CT images.

We include the conditional generation results from the baseline model without considering the fracture loss in Fig. 10. Without using the fracture loss, the baseline DDPMs are not able to generate fractures that are consistent with those in the reference micro-CTs. This demonstrates the significant improvement achieved by adding the fracture loss.

The autoencoder were trained with with a discriminator, Kullback–Leibler (KL) divergence, and perceptual losses. Figs. 11 and 12 demonstrates that the autoencoder is able to reconstruct the original micro-CT well given a certain set of noise. The SiT-predicted micro-CT images demonstrate superior accuracy compared to DDPMs predictions as evidenced by both SSIM (avg=0.41) and PSNR (avg=20.71) metrics illustrated in Figs. 6 and 7. Notably, these results were achieved despite SiT being trained for fewer than 100 epochs, which is significantly less than the 200 to 300 epochs allocated to fully trained DDPMs. Furthermore, SiT's accuracy was evaluated on full-resolution micro-CT images rather than the resized versions used for DDPMs. This enhances physical fidelity. Qualitatively, as shown in Figs. 13 and 14, SiT predictions exhibit markedly better preservation of fine structural details including small black grains, micro-fractures, and intricate grain textures. These

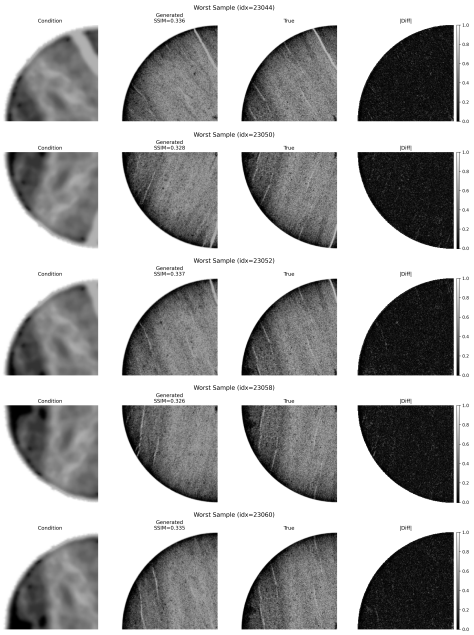

Figure 9. The 5 samples predicted by the DDPMs with fracture loss that have the worst SSIM among the 60 test images. Each row (from left to right) contains the conditioning CT, the generated micro-CT, the true micro-CT, and the absolute pixel-wise prediction error.

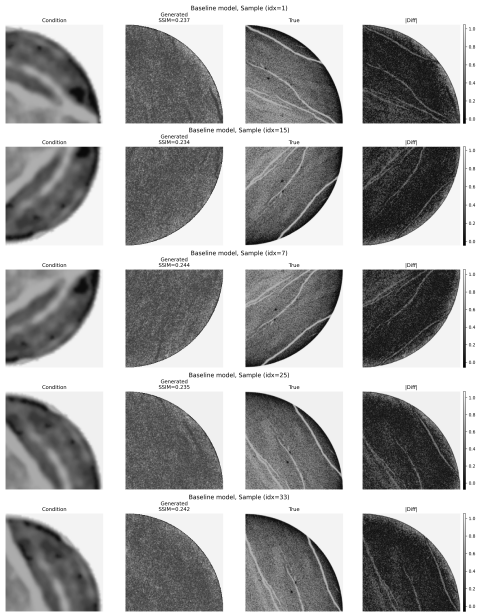

Figure 10. The 5 samples predicted by the DDPMs without fracture loss that achieved the highest SSIM among the 60 test images. Each row (from left to right) contains the conditioning CT, the generated micro-CT, the true micro-CT, and the absolute pixel-wise prediction error.

findings suggest that SiT not only converges faster but also preserves more physically meaningful features compared to DDPMs. Future work will include direct performance comparisons between SiT and latent diffusion models to provide a more valid assessment.

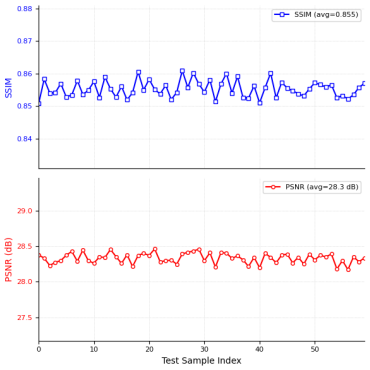

Figure 11. The structural similarity index measure and peak signal-to-noise ratio of 60 images from the test dataset decoded by the autoencoder.

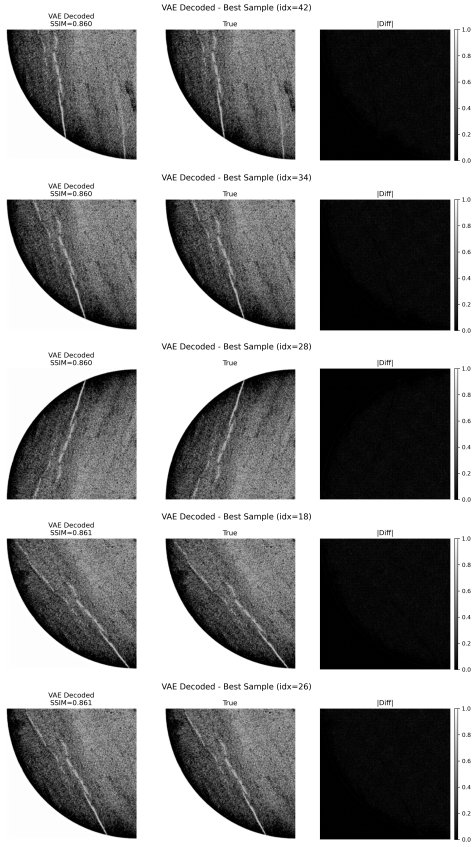

Figure 12. The 5 samples predicted by the autoencoder that achieved the highest SSIM among the 60 test images. Each row (from left to right) contains the decoded micro-CT, the true micro-CT, and the absolute pixel-wise prediction error.

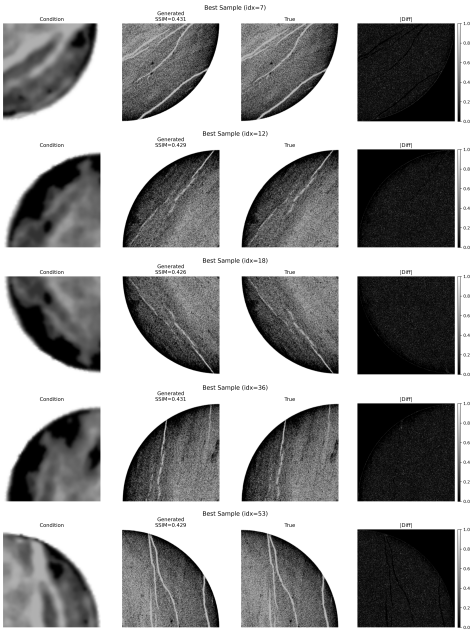

Figure 13. The 5 samples predicted by the SiT that achieved the highest SSIM among the 60 test images. Each row (from left to right) contains the decoded micro-CT, the true micro-CT, and the absolute pixel-wise prediction error.

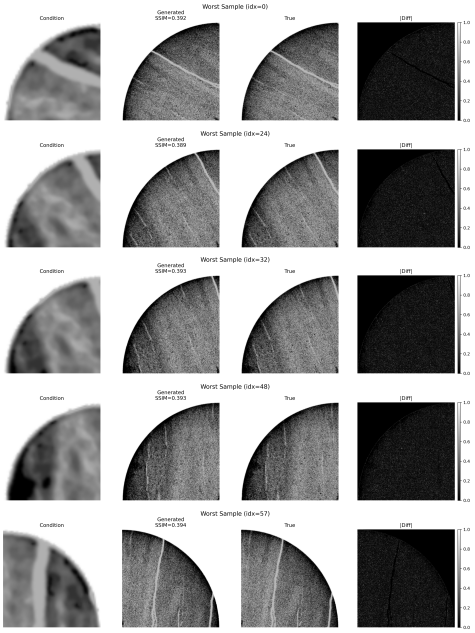

Figure 14. The 5 samples predicted by the SiT that achieved the lowest SSIM among the 60 test images. Each row (from left to right) contains the decoded micro-CT, the true micro-CT, and the absolute pixel-wise prediction error.

7. Discussion

In this study, we treat the image super-resolution as a conditional image generation task. We mainly explored two models: Denoising diffusion probabilistic models (DDPMs) and scalable interpolant transformer (SiT). Owing to time constraints, we did not investigate the latent diffusion model, which employs a U-Net architecture as its backbone. Instead, we resized the micro-CT image from (512,512) to (256,256) and trained DDMPs directly on the resized images. For the SiT training, we first trained an auto-encoder with a discriminator, KL divergence, and perceptual losses to reduce the dimension of micro-CT images from (1,512,512) to (8,64,64). Then, the SiT training is performed on the encoded latent space with the CT images of size (1,64,64) as another concatenated conditioning channel. Overall, we have achieved at least satisfactory performance on all models examined.

Inference on 60 reduced-size images using the trained DDPMs takes approximately 9 minutes with 800 timesteps. In contrast, inference with the autoencoder and SiT models on 60 full-size images requires only around 6 minutes. This performance difference is rooted in the fundamental architectural distinctions between these generative approaches. DDPMs operate through an iterative Markov chain of denoising steps, and they require hundreds of sequential forward passes through the network to gradually convert noise into image structure. This inherently sequential nature creates a computational bottleneck that scales linearly with the number of diffusion steps.

In addition, the SiT models outperform DDPMs in both perceptual quality and numerical accuracy. For DDPMs, the inclusion of fracture loss is critical because it helps direct the model's attention to essential features. Without this guidance, DDPMs often fail to converge to the ground truth. These observations suggest that SiT training is inherently more stable and better guided than that of DDPMs. The transformer-based architecture of SiT leverages self-attention mechanisms that capture global dependencies across the entire spatial domain in a single operation. This parallel processing of spatial relationships allows SiT to model complex inter-dependencies more effectively than the step-by-step refinement approach of DDPMs. The SiT architecture's theoretical advantage stems from its formulation as a continuous normalizing flow in the latent space. By directly modeling velocity fields rather than noise residuals, SiT learns probability flow ODEs that provide more direct paths through the latent space. This learning reduces the variance in gradient estimates during training compared to the noise prediction objective in DDPMs. Because the model learns a deterministic transformation rather than relying on a stochastic process with potentially higher variance, the flow-based formulation also enables more stable convergence properties. Combining with the transformer's capacity to capture long-range dependencies, the flow-based strategy leads to the observed improvements in both generation quality and computational efficiency.

8. Conclusion and Future Work

This study addresses the problem of rock image super-resolution by framing it as a conditional generation task. We evaluate two representative generative models, DDPM and SiT, for synthesizing high-resolution micro-CT images from low-resolution CT inputs. Both models demonstrate the ability to recover fine structural details and generate realistic micro-CT representations conditioned on coarse-resolution data.

Among the two models, SiT demonstrates superior performance, not only through visual comparison but also based on quantitative metrics such as SSIM and PSNR. While DDPM provides satisfactory results, SiT offers additional advantages in computational efficiency by operating in a lower-dimensional latent space, which reduces the cost of the iterative generation process. These results highlight the effectiveness of deep generative models in bridging the resolution gap between CT and micro-CT imagery, providing a data-driven approach for enhancing geological interpretation.

Future work will focus on extending this framework in several directions. First, we plan to implement DiT and other transformer architectures to compare both generation quality and computational performance. Second, we aim to integrate end-to-end learning with visual representations derived from DINO, which may lead to more robust feature learning. Third, we want to implement latent diffusion model to give a more valid comparison between SiT and DDPMs. Finally, we will expand the dataset to include various rock types, such as shale and mudstone, with the goal of training a single model that generalizes across lithologies.

9. Contributions & Acknowledgements

Zitong Huang and Minghui Xu equally contributed to this project. We thank the CS231n teaching team, especially Serena Zhang for her supports. The entirety of this project is specifically done for 231n. We obtained the base codes of DDPMs from: https://huggingface.co/blog/annotated-diffusion and the base codes of SiT from: https://github.com/willisma/SiT

References

[1] M. S. Albergo, N. M. Boffi, and E. Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions, 2023. 2

[2] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis, 2021. 1

[3] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2, 3

[4] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling, 2023. 2

[5] C. Liu, Y. Liu, L. Shan, S. V. Chilukoti, and X. Hei. Enhancing unsupervised rock ct image super-resolution with non-local attention. *Geoenergy Science and Engineering*, 238:212912, 05 2024. 2

[6] X. Liu, C. Gong, and Q. Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. 2

[7] N. Ma, M. Goldstein, M. S. Albergo, N. M. Boffi, E. Vanden-Eijnden, and S. Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers, 2024. 1

[8] M. P. Murugesu, V. Krishnan, and A. R. Kovscek. Enhancing prediction of fluid-saturated fracture characteristics using deep learning super resolution. *Applied Computing and Geosciences*, 24:100208, 2024. 1, 2, 4

[9] W. Peebles and S. Xie. Scalable diffusion models with transformers, 2023. 1

[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2022. 3

[11] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015. 3

[12] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution, 2020. 2

[13] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations, 2021. 2

[14] C. You, G. Li, Y. Zhang, X. Zhang, H. Shan, M. Li, S. Ju, Z. Zhao, Z. Zhang, W. Cong, M. Vannier, P. Saha, E. Hoffman, and G. Wang. Ct super-resolution gan constrained by the identical, residual, and cycle learning ensemble (gan-circle). *IEEE Transactions on Medical Imaging*, PP:1–1, 06 2019. 2

[15] M. Zhaoyang, S. Sun, B. Yan, H. Kwak, and J. Gao. Enhancing the resolution of micro-ct images of rock samples via unsupervised machine learning based on a diffusion model. 10 2023. 2

[16] S. Zhou, L. Yu, and M. Jin. Texture transformer super-resolution for low-dose computed tomography. *Biomedical physics engineering express*, 8, 11 2022. 2