Diffusion-Based Super-Resolution of Micro–CT to SEM for Porous Media

Hanqi Li
Stanford University
450 Jane Stanford Way, Stanford, CA 94305
hanqili7@stanford.edu

Abstract

This project aims to generates high-fidelity scanning-electron-microscope (SEM) detail from lower-resolution micro-CT slices of rock. Using two co-registered images from a public North Sea sandstone dataset, we build a large paired training set via image augmentation. Our solution is a conditional denoising-diffusion probabilistic model (DDPM) in which only the SEM channel is noised, while the micro-CT slice remains as a fixed condition. We test our approach against a standard conditional GAN baseline.

1. Introduction

High-resolution imaging of geological materials is crucial for understanding pore-scale features that govern fluid flow, mechanical strength, and reactive transport in subsurface formations. While SEM can resolve sub-micron textures and fine-grained mineral phases, obtaining large-area or volumetric datasets is prohibitively time-consuming. Sample preparation—cutting, polishing, coating, and mounting—can take hours per sample, and acquisition requires careful tuning of beam parameters, working distance adjustments, and iterative focusing. Consequently, mapping an entire surface at SEM resolution can demand days of continuous instrument operation, significantly slowing both exploratory studies and high-throughput analysis.

Micro–computed tomography (micro-CT) offers non-destructive 3-D imaging at voxel sizes of a few micrometers, enabling rapid scanning of centimeter-scale cores with minimal preparation. However, its spatial resolution is insufficient to resolve sub-micron structural details—such as nanopores, grain boundaries, and micro-cracks—that control processes like capillary trapping, mineral dissolution, and cement phase redistribution. Researchers therefore face a trade-off between field of view (with micro-CT) and fine-scale detail (with SEM), leaving a gap in comprehensive volumetric characterization.

Image super-resolution (SR)[11] seeks a mapping

$$\mathcal{F} : \mathbf{X}_{\mathrm{low}} \longrightarrow \mathbf{X}_{\mathrm{high}}$$

that reconstructs plausible high–resolution content from coarse measurements. In our context, SR enables generation of SEM-quality textures from micro-CT slices without the need for exhaustive SEM acquisition. This accelerated approach reduces instrument time, minimizes repetitive sample handling and tuning, and facilitates high-throughput workflows where large sample sets must be characterized consistently.

Classical SR techniques—interpolation, sparse coding, or shallow CNNs—struggle with the highly non-linear relationship between X-ray attenuation and electron backscatter. Denoising-diffusion probabilistic models (DDPMs) offer a principled alternative: by learning to reverse a gradual noising process, they excel at synthesizing high-frequency detail while preserving global conditioning cues. We therefore cast micro-CT → SEM SR as a conditional diffusion task, injecting noise only into the SEM channel and guiding the reverse process with the aligned micro-CT slice.

Problem statement

The input to our algorithm during training is a paired micro-CT and SEM slice. We then use a denoising-diffusion probabilistic model to output a predicted SEM-quality image conditioned on the micro-CT input.

Given paired images

$$x^{\mathrm{MCT}},\, x^{\mathrm{SEM}} \;\in\; \mathbb{R}^{1\times512\times512},$$

where x^{MCT} is a 512×512 micro–CT slice and x^{SEM} the corresponding SEM slice, learn a conditional generative model

$$p_\theta\big(x^{\mathrm{SEM}} \mid x^{\mathrm{MCT}}\big)$$

that produces SEM-quality textures consistent with unseen micro–CT inputs. Generated outputs are evaluated against the ground-truth images using PSNR, SSIM, and additional similarity metrics to assess accuracy.

2. Related Work

Below, we review existing methods for image super-resolution (SR) in the context of micro-CT and SEM imaging, grouping approaches into CNN-based methods, GAN-based frameworks, and cross-modal applications specific to geological materials. We highlight strengths and weaknesses of representative works and identify gaps that motivate our diffusion-based approach.

2.1. CNN-Based Super-Resolution for Micro-CT and Digital Rocks

With the development of deep learning, convolutional neural networks (CNNs) became one of the dominant approach for SR tasks. The seminal Super-Resolution Convolutional Neural Network (SRCNN) proposed by Dong et al. [4] demonstrates that an end-to-end CNN can learn the mapping from low- to high-resolution images, outperforming traditional sparse-coding methods on natural images [4, 10]. Inspired by SRCNN, Zhang et al. [13] developed a multi-scale fusion residual U-Net (MS-ResUnet) to enhance rock micro-CT images, achieving significant gains over bicubic interpolation by leveraging hierarchical feature learning [13, 2]. These CNN-based methods excel at capturing local textures through hierarchical feature extraction and deliver substantial quantitative improvements over classical baselines. However, they often struggle to maintain global consistency across large fields of view, producing blocky or tiled artifacts when applied to regions outside their training distribution. Furthermore, their reliance on substantial paired training data limits robustness when applied to novel lithologies or datasets with limited co-registered micro-CT/SEM pairs.

2.2. GAN-Based Frameworks for 3D Micro-CT Super-Resolution

Generative adversarial networks (GANs) have been extended to volumetric micro-CT SR to enhance perceptual quality and segmentation accuracy. For example, Ugolkov et al. [14] proposed a memory-efficient 3D octree-based WGAN-GP that achieves 16× resolution enhancement and corrects segmentation inaccuracies caused by overlapping X-ray attenuation in micro-CT measurements [14, 15]. By embedding an octree structure, this approach overcomes the memory bottleneck inherent in 3D convolutions, enabling super-resolution from 7 µm to 0.44 µm per voxel. Prior to this, the same group introduced a 3D DC WGAN-GP for eightfold SR on Berea sandstone, which similarly improved segmentation accuracy for minerals and pore space but still required large unpaired training sets to cover diverse lithologies [15]. Wang et al. [3] developed EDSR-GAN—an enhanced deep SR GAN for digital rock images—that demonstrated a 50–70% reduction in relative error compared to interpolation and recovered sub-resolution features such as dissolved minerals and fractures [3, 8]. Together, these GAN-based frameworks produce perceptually realistic outputs and sharpen fine details, aiding downstream tasks like segmentation. However, their high memory requirements and architectural complexity make them challenging to train and deploy, and they depend heavily on sizable, accurately labeled paired datasets. Additionally, GANs may suffer from training instabilities such as mode collapse, which limits their reliability for large-scale volumetric applications.

2.3. Cross-Modal and SEM-Specific Super-Resolution

Cross-modal SR methods synthesize SEM-quality textures from micro-CT inputs, demonstrating that inferred sub-resolution features such as micro-porosity align with actual pore-scale geometry. For example, Wang et al. [3] validated EDSRGAN outputs against SEM images, showing substantial reductions in relative error and realistic recovery of fine details like dissolved minerals and fractures [3, 8]. Similarly, Armstrong et al. [12] applied CNNs to denoise and deblur micro-CT data to improve segmentation accuracy, though they did not explicitly target SEM-like texture synthesis [12, 2]. In parallel, SEM-specific SR approaches in materials science—for instance, Berzins et al. [5] super-resolved SEM images of Li-ion cathode materials to enhance crack segmentation—highlight the downstream benefits of improved 2D SEM quality for defect detection [5]. While these techniques underscore the potential of SR for both geological and materials applications, they remain constrained by labor-intensive SEM acquisition for each region of interest and are typically demonstrated on limited 2D or small 3D volumes, limiting throughput for large-scale studies.

2.4. Summary

CNN-based methods (e.g., SRCNN [4], MS-ResUnet [13], EDSR [3, 8]) leverage hierarchical feature extraction to achieve substantial quantitative improvements over interpolation, but they often exhibit blocky artifacts when applied beyond their training distribution and require large paired datasets to generalize across varied lithologies. GAN-based frameworks (e.g., EDSRGAN [3], octree-WGAN [14, 15]) generate perceptually realistic volumetric outputs and improve segmentation accuracy, yet their high memory demands, architectural complexity, and potential for training instability limit scalability. Cross-modal approaches that synthesize SEM-like textures from micro-CT inputs demonstrate that inferred textures can faithfully reflect pore-scale geometry [3, 12], but they remain constrained by time consuming co-registration and extensive SEM imaging. SEM-specific SR in materials science (e.g., Berzins et al. [5]) shows clear benefits for downstream 2D analyses such as defect detection, yet these methods do not extend to full 3D volumes and still depend on time-consuming SEM acquisition. Currently, most practitioners continue to perform high-resolution SEM imaging manually, as automated SR solutions are not yet fully validated in industrial or research pipelines.

3. Methods

To our knowledge, no prior work has applied denoising-diffusion probabilistic models (DDPMs) to the micro-CT → SEM super-resolution (SR) problem. Formally, let $x^{\mathrm{mct}} \in \mathbb{R}^{1\times H\times W}$ denote a low-resolution micro-CT slice and $x^{\mathrm{sem}} \in \mathbb{R}^{1\times H\times W}$ its corresponding high-resolution SEM slice. Our goal is to train a model $\mathcal{F} : \big(x^{\mathrm{mct}},\, z\big) \mapsto \hat{x}^{\mathrm{sem}}$ that produces a plausible SEM-quality image $\hat{x}^{\mathrm{sem}}$ given x^{mct} (and optionally noise z). As a reference, we use a conditional GAN (cGAN) architecture—widely adopted in image-to-image translation—where the generator G is conditioned on the micro-CT input and trained to fool a discriminator D that receives both the generated output and the original low-resolution slice. Representative cGAN frameworks (e.g., Pix2Pix [7], ESRGAN [16]) can yield sharp textures but often suffer from mode collapse, training instability, and the need for carefully balanced adversarial and reconstruction losses to achieve both fidelity and realism.

Denoising-diffusion probabilistic models (DDPMs) [6] sidestep the adversarial min–max game by casting super-resolution as maximum-likelihood estimation in a latent noise space: a Gaussian corruption process iteratively destroys high-frequency content in the SEM channel, a neural network learns to predict the added noise at each timestep, and ancestral denoising inverts this chain while concatenating the clean micro-CT slice for perfect conditioning. This single reconstruction objective yields stable optimization, avoids mode collapse, and has already pushed the state of the art in image SR (e.g., SR3, Palette) beyond cGAN baselines in both perceptual scores and Fréchet Inception Distance. In the following subsections, we will elaborate on the formulations of both the DDPM and the cGAN for micro-CT → SEM super-resolution.

3.1. Conditional cGAN Baseline

A conditional cGAN treats image SR as an adversarial game between a generator G and a discriminator D. The generator G_ϕ takes the micro-CT input x^{mct} (normalized to $[-1, 1]$) and outputs a super-resolved image $\hat{x}^{\mathrm{sem}} = G_\phi(x^{\mathrm{mct}})$. The discriminator D_ψ receives a pair $\big(x^{\mathrm{mct}},\, x\big)$, where x is either a real SEM slice x^{sem} or a generated output $\hat{x}^{\mathrm{sem}}$, and predicts $D_\psi\big(x^{\mathrm{ct}},\, x\big) \in [0, 1]$, indicating whether x is real.

Generator architecture. The generator G_ϕ is implemented as a lightweight U-Net, denoted `UNet1`, that maps a single-channel micro-CT input $x^{\mathrm{mct}} \in \mathbb{R}^{1\times 512\times 512}$ to a single-channel SEM output $\hat{x}^{\mathrm{sem}}$. First, the input passes through an encoding convolution $\mathrm{Conv2d}(1, 64, 4, 2, 1)$ with InstanceNorm2d and LeakyReLU, reducing the resolution from 512×512 to 256×256. Next, a “mid” convolution $\mathrm{Conv2d}(64, 128, 4, 1, 1)$ (stride=1, padding=1) further processes the 256×256 features without spatial downsampling. In the decoder, a transpose convolution $\mathrm{ConvTranspose2d}(128, 64, 4, 2, 1)$ with InstanceNorm2d and ReLU upsamples back to 512×512. Finally, the decoder output is bilinearly interpolated to match the original spatial dimensions if necessary, concatenated with the original x^{mct} (skip connection), and passed through a 1×1 convolution that reduces $64 + 1$ channels to 1, followed by a Tanh activation to produce $\hat{x}^{\mathrm{sem}} \in [-1, 1]$.

Discriminator architecture. The discriminator D_ψ is a PatchGAN (`PatchD`) that takes the concatenated pair $\big[x^{\mathrm{mct}},\, x\big] \in \mathbb{R}^{2\times 512\times 512}$ (where x is either a real SEM slice or a generated output) and outputs a 63×63 feature map of “real”/“fake” scores. Concretely, four successive blocks apply $\mathrm{Conv2d}$ layers with kernel size 4 and stride 2 (except the fourth block uses stride=1), each followed by InstanceNorm2d and LeakyReLU(0.2). The channel progression is 2 → 64 → 128 → 256 → 512, and all convolutional layers are wrapped with spectral normalization. A final 1×1 convolution maps 512 channels to 1, yielding a patch-score map.

Loss functions. We train G_ϕ and D_ψ with hinge adversarial losses and an ℓ_1 reconstruction term. For a minibatch of real pairs $\big(x^{\mathrm{mct}},\, x^{\mathrm{sem}}\big)$ and generated outputs $\hat{x}^{\mathrm{sem}} = G_\phi\big(x^{\mathrm{mct}}\big)$, the discriminator loss can be written in a split form to avoid overflow:

$$
\begin{split}
\mathcal{L}_D = \mathbb{E}_{(x^{\mathrm{mct}},\, x^{\mathrm{sem}})}\Big[\max\big(0,\ 1\ -\ D_\psi\big(x^{\mathrm{mct}},\, x^{\mathrm{sem}}\big)\big)\Big] \\
+ \mathbb{E}_{x^{\mathrm{mct}}}\Big[\max\big(0,\ 1\ +\ D_\psi\big(x^{\mathrm{mct}},\, G_\phi(x^{\mathrm{mct}})\big)\big)\Big].
\end{split}
$$

The generator minimizes a combination of the adversarial hinge loss and an ℓ_1 reconstruction loss, also formatted in a split environment:

$$
\begin{split}
\mathcal{L}_G = -\,\mathbb{E}_{x^{\mathrm{mct}}}\Big[D_\psi\big(x^{\mathrm{mct}},\, G_\phi(x^{\mathrm{mct}})\big)\Big] \\
+ \lambda_{\ell_1}\, \mathbb{E}_{(x^{\mathrm{mct}},\, x^{\mathrm{sem}})}\Big[\big\|x^{\mathrm{sem}} - G_\phi(x^{\mathrm{mct}})\big\|_1\Big],
\end{split}
$$

where $\lambda_{\ell_1} = 100$. Both G_ϕ and D_ψ are optimized with Adam ($\mathrm{lr} = 2 \times 10^{-4}, \beta_1 = 0.5, \beta_2 = 0.999$).

3.2. Conditional DDPM Implementation

In our super-resolution task, each training example consists of a perfectly aligned pair $(x^{\mathrm{mct}}, x^{\mathrm{sem}})$, where $x^{\mathrm{mct}} \in \mathbb{R}^{1\times H\times W}$ is a low-resolution micro-CT slice and $x^{\mathrm{sem}} \in \mathbb{R}^{1\times H\times W}$ is its corresponding high-resolution SEM image. We construct a two-channel tensor

$$x^{(0)} = \begin{bmatrix} x_0^{\mathrm{sem}} \\ x^{\mathrm{mct}} \end{bmatrix} \in \mathbb{R}^{2\times H\times W},$$

in which only the SEM channel is corrupted by Gaussian noise during the forward diffusion process, while the micro-CT channel remains entirely noise-free. This conditional setup ensures that the U-Net denoiser learns to leverage the low resolution structure present in x^{mct} to “fill in” fine-scale SEM details rather than synthesizing the entire image from noise.

At each timestep $t = 1, \ldots, T$, we sample $\varepsilon \sim \mathcal{N}(0, I)$ (gaussian noise) of shape $(2, H, W)$ and zero out its second channel so that

$$\varepsilon = \begin{bmatrix} \varepsilon^{\mathrm{sem}} \\ \mathbf{0} \end{bmatrix}.$$

The SEM channel evolves according to

$$x_t^{\mathrm{sem}} = \sqrt{\alpha_t}\, x_{t-1}^{\mathrm{sem}} + \sqrt{1-\alpha_t}\, \varepsilon^{\mathrm{sem}}, \qquad x_t^{\mathrm{mct}} = x_0^{\mathrm{mct}}.$$

Equivalently, one may write the marginal forward-diffusion distribution:

$$\begin{aligned} q\big(x^{(t)} \mid x^{(0)}\big) &= q\big(x_t^{\mathrm{sem}} \mid x_0^{\mathrm{sem}}\big)\, \delta\big(x_t^{\mathrm{mct}} - x^{\mathrm{mct}}\big) \\ &= \mathcal{N}\Big(\sqrt{\bar\alpha_t}\, x_0^{\mathrm{sem}},\, (1-\bar\alpha_t)\, I\Big)\, \delta\big(x_t^{\mathrm{mct}} - x^{\mathrm{mct}}\big), \end{aligned} \tag{1}$$

$$\bar\alpha_t = \prod_{s=1}^{t} \alpha_s.$$

Thus, at every t, the second channel remains exactly x^{mct}, and only the first channel x_t^{sem} is a noisy version of the original SEM.

To reverse this process, we employ a U-Net denoiser ε_θ which is explicitly modified to accept two input channels: the noisy SEM slice x_t^{sem} and the clean micro-CT slice x^{mct}. In implementation, the first convolution of the U-Net is adjusted from $1 \to d$ to $2 \to d$ channels, and every residual and attention block propagates the micro-CT feature map alongside the SEM feature map. During training, we randomly draw $t \sim \mathrm{Uniform}(\{1, \ldots, T\})$ and sample $\varepsilon^{\mathrm{sem}} \sim \mathcal{N}(0, I)$ to compute

$$x_t^{\mathrm{sem}} = \sqrt{\bar\alpha_t}\, x_0^{\mathrm{sem}} + \sqrt{1-\bar\alpha_t}\, \varepsilon^{\mathrm{sem}}, \qquad x^{(t)} = \begin{bmatrix} x_t^{\mathrm{sem}} \\ x^{\mathrm{mct}} \end{bmatrix},$$

and then minimize the smooth-ℓ_1 loss

$$\mathcal{L}_{\mathrm{DDPM}} = \mathbb{E}_{x_0^{\mathrm{sem}},\, x^{\mathrm{mct}},\, t,\, \varepsilon^{\mathrm{sem}}} \left\| \varepsilon^{\mathrm{sem}} - \varepsilon_\theta\big([x_t^{\mathrm{sem}},\, x^{\mathrm{mct}}],\, t\big) \right\|_1 .$$

We optimize θ with AdamW and clip gradients to a maximum norm of 0.7, reducing the learning rate on plateau of the validation loss. Since no noise is ever added to the micro-CT channel, its features remain intact as a guide in every U-Net block.

At inference, we initialize $x_T^{\mathrm{sem}} \sim \mathcal{N}(0, I)$ and iterate $t = T, T-1, \ldots, 1$. Given the current noisy state $x^{(t)} = [x_t^{\mathrm{sem}},\, x^{\mathrm{mct}}]$, we predict

$$\hat\varepsilon_\theta = \varepsilon_\theta\big([x_t^{\mathrm{sem}},\, x^{\mathrm{mct}}],\, t\big), \qquad \mu_\theta = \frac{1}{\sqrt{\alpha_t}} \left(x_t^{\mathrm{sem}} - \frac{1-\alpha_t}{\sqrt{1-\bar\alpha_t}}\, \hat\varepsilon_\theta \right),$$

and sample

$$x_{t-1}^{\mathrm{sem}} = \mu_\theta + \sqrt{\beta_t}\, \eta_t, \qquad \eta_t \sim \mathcal{N}(0, I),$$

while concatenating x^{mct} to form

$$[x_{t-1}^{\mathrm{sem}},\, x^{\mathrm{mct}}].$$

After T denoising steps, x_0^{sem} is our final SEM-quality reconstruction $\hat{x}^{\mathrm{sem}}$.

Compared to standard, unconditional DDPM code (such as the Hugging Face “annotated diffusion” notebook), our implementation differs in that:

1. we use a custom `rockDataset` to load paired micro-CT/SEM tiles and apply identical augmentations to both channels,

2. we zero out any noise in the micro-CT channel so only the SEM is corrupted, and

3. we modify the U-Net’s input to two channels and propagate the clean micro-CT features throughout every convolutional block.

By conditioning in this manner—always re-concatenating x^{mct} when denoising—the model learns to reconstruct high-frequency SEM details guided by low-frequency micro-CT structure rather than generating from pure noise.

4. Dataset

4.1. Source data

The dataset used in this project is pixel–registered micro-CT and SEM pair for real rock porous media from the *North Sea Sandstone* release on Digital Porous Media.[1] It contains two aligned 6100 × 6100 images: one micro–CT and one SEM, however, one pair is insufficient to train a diffusion model, so we built an aggressive tiling plus augmentation pipeline to synthesise a dataset of 36 785 perfectly aligned image pairs.

[1] `https://digitalporousmedia.org/published-datasets/tapis/projects/drp.project.published/drp.project.published.DRP-251/`

4.2. Image augmentation

The mCT slice was adjusted brightness and contrast in FIJI/ImageJ to enhance pore–grain contrast, then exported as 8-bit PNG. The SEM slice was likewise converted to 8-bit so that both images share the same dynamic range. No further registration was required.

We slide a 512×512 window across each 6100×6100 image with a stride of $128\,\mathrm{px}$ (75 % overlap). Each axis therefore yields

$$n_x = n_y = \left\lfloor \frac{6100 - 512}{128} \right\rfloor + 1 = 44,$$
$$1936 = 44 \times 44 \quad \text{base tiles per image.}$$

Applying this grid to both mCT and SEM produces 1936 aligned 512×512 image pairs per image, to further grow the dataset without padding artifacts, we use a three-step rotation-safe augmentation:

1. **Pad-free crop:** extract a centered patch of size $512\sqrt{2}$ so that any in-plane rotation remains fully inside it.
2. **In-plane rotations:** rotate this patch by $\theta \in \{0°, 20°, \ldots, 340°\}$.
3. **Center crop:** from each rotated view, center-crop back to 512×512.

The process yields 19 views per base tile. In practice the script is executed on all original tiles extracted from the mCT/SEM pair, giving

$$\underbrace{1\,936}_{\text{base tiles}} \times 19 = 36\,784$$

aligned training pairs. This volume of data ensures that we have sufficient samples for training, validation, and testing.

Figure 1 shows one base tile (unrotated) and two augmented (rotated) examples for both micro-CT (top row) and SEM (bottom row). Note how the pore and grain geometry in the micro-CT slice corresponds to fine-grained textures in the SEM slice, even after rotation.

After gathering enough data, we randomly assign 24 000 pairs to the training set, 4 800 to the validation set, and keep the remainder for testing purposes. No tile from a given location appears in more than one split, guaranteeing a clean separation between train and validation data.

5. Results

In this section, we present both quantitative and qualitative evaluations of our conditional GAN (cGAN) baseline and our proposed denoising-diffusion probabilistic model (DDPM) for micro-CT → SEM super-resolution. We first define the primary metrics (SSIM and PSNR) used to compare generated SEM images against ground truth. Next, we summarize our training hyperparameters and experimental setup, and then present quantitative results followed by qualitative examples illustrating typical successes and failure modes for both methods. Finally, we discuss overfitting considerations and practical trade-offs between the two approaches.

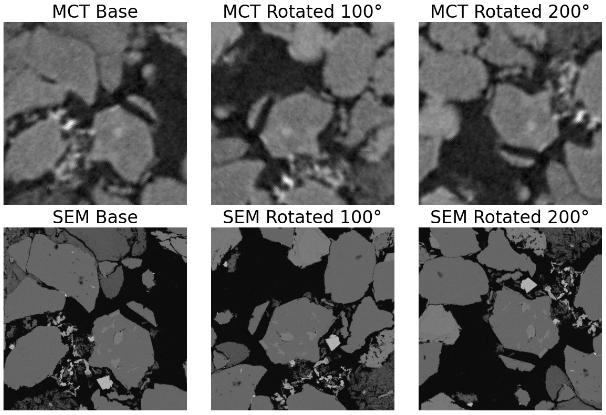

Figure 1. Example micro-CT and SEM pair tiles from our dataset

5.1. Evaluation Metrics

To quantitatively assess the fidelity of generated SEM images $\hat{x}^{\mathrm{sem}}$ relative to ground-truth SEM slices x^{sem}, we employ two standard image-quality metrics: the structural similarity index measure (SSIM) and the peak signal-to-noise ratio (PSNR). Let $\hat{x}$ and x denote two single-channel images of size $H \times W$.

Structural Similarity Index Measure (SSIM). SSIM measures perceptual similarity by comparing local luminance, contrast, and structural information [1]. For any local patch x_i and $\hat{x}_i$, SSIM is defined as

$$\mathrm{SSIM}(x_i, \hat{x}_i) \;=\; \frac{(2\mu_x\mu_{\hat{x}} + C_1)\,(2\sigma_{x\hat{x}} + C_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + C_1)\,(\sigma_x^2 + \sigma_{\hat{x}}^2 + C_2)},$$

where $\mu_x, \mu_{\hat{x}}, \sigma_x^2, \sigma_{\hat{x}}^2$ are the means and variances over the patch, $\sigma_{x\hat{x}}$ is the covariance, and C_1, C_2 are small constants for numerical stability. The overall SSIM index between full-resolution images is the average of SSIM over a sliding window. Values range from 0 to 1, with 1 indicating perfect structural agreement.

Peak Signal-to-Noise Ratio (PSNR). PSNR measures pixel-wise fidelity by comparing the mean-squared error (MSE) to the dynamic range of pixel intensities[9]. Let

$$\mathrm{MSE}(x, \hat{x}) \;=\; \frac{1}{HW} \sum_{p=1}^{H} \sum_{q=1}^{W} \bigl(x_{p,q} - \hat{x}_{p,q}\bigr)^2,$$

where $x, \hat{x} \in [0,1]^{H\times W}$. Then

$$\mathrm{PSNR}(x,\hat{x}) \;=\; 10\log_{10}\!\Big(\frac{L^2}{\mathrm{MSE}(x,\hat{x})}\Big), \quad L=1.$$

Higher PSNR indicates smaller reconstruction error.

5.2. Quantitative Results

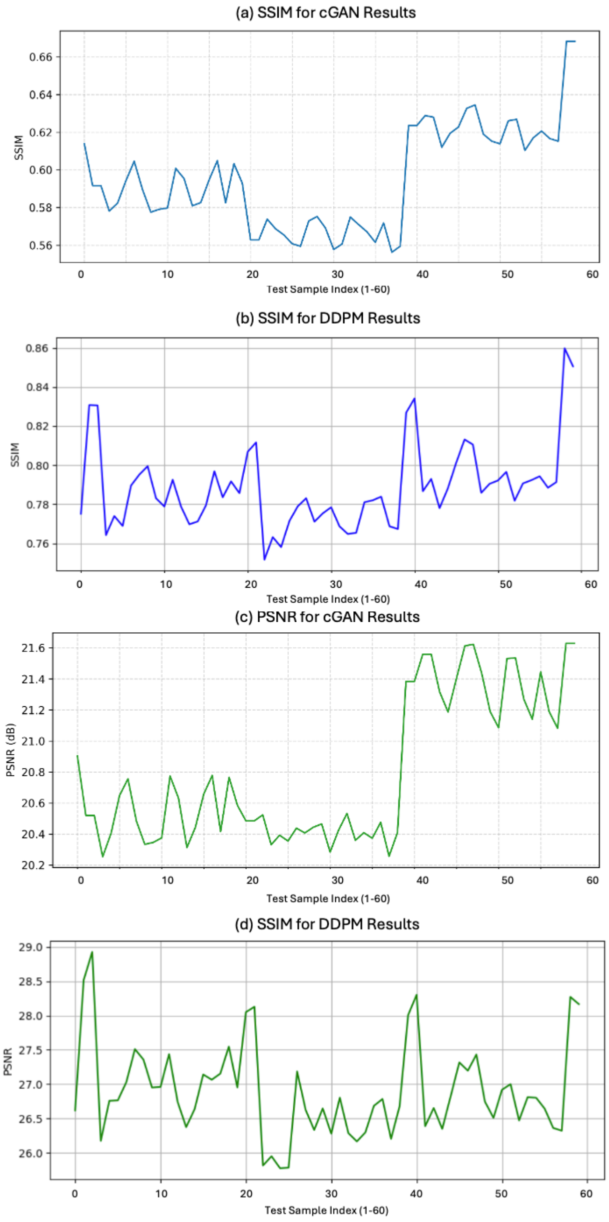

Figure 2. SSIM and PSNR curves for 60 held-out test examples. Panels (a) and (c) display cGAN baseline performance; panels (b) and (d) display DDPM performance.

All reported results use the single checkpoint from each model that achieved the lowest validation loss. We then evaluated both models on a subset of 60 SEM–micro-CT image pairs that were never seen during training. Figure 2 plots SSIM and PSNR curves over the same 60 test examples for both the cGAN baseline and our DDPM.

The SSIM trajectory for the cGAN (Figure 2(a)) ranges from approximately 0.55 to 0.67, with an overall average of 0.60. Lower SSIM values appear early in the sequence—around indices 1–20—indicating that these particular micro-CT geometries produce more challenging textures for the U-Net generator. As the index increases past 40, SSIM steadily improves; for example, test sample 59 reaches nearly 0.67, suggesting that the cGAN captures larger, smoother grain boundaries more faithfully. The corresponding PSNR curve (Figure 2(c)) similarly begins near 20.2 dB (indices 1–5), dips to 20.0 dB around index 20, and then climbs to a maximum of 21.6 dB at index 59. This trend implies that pixel-wise reconstruction error decreases on those "easier" samples with more uniform contrast.

In contrast, the DDPM's SSIM (Figure 2(b)) lies between 0.75 and 0.86, with a mean of 0.80—substantially higher than the cGAN across every index. Although occasional dips (e.g. index 20 at 0.75) occur when the SEM contains highly irregular pore networks, the DDPM consistently recovers structural details that the cGAN misses. Notably, index 59 achieves SSIM 0.86, indicating almost perfect agreement with the ground-truth SEM in regions of minimal high-frequency noise. The DDPM's PSNR (Figure 2(d)) spans 25.7–29.0 dB (mean 27.0 dB), clearly outperforming the cGAN's 20.0–21.6 dB range. Peaks near 29 dB correspond to samples whose SEMs are dominated by smooth mineral phases, which the diffusion model can reconstruct with very low pixel-wise error.

Table 1 reports the average SSIM and PSNR over the full set of 7 985 test pairs. On average, the cGAN achieves SSIM 0.595 and PSNR 20.52 dB, whereas our DDPM obtains SSIM 0.789 and PSNR 26.91 dB. These aggregate metrics confirm that selecting the best validation checkpoint for each architecture yields a substantial performance gap: the DDPM outperforms the cGAN baseline by nearly 0.21 SSIM points and 6.56 dB in PSNR on unseen rock-texture images.

Table 1. Average SSIM and PSNR on the full unseen test set (7 985 pairs).

Model	**Mean SSIM**	**Mean PSNR (dB)**
cGAN (best-val)	0.595	20.52
DDPM (best-val)	0.789	26.91

5.3. Qualitative Results

Figure 3 presents three randomly selected test examples (rows) that are identical for both the cGAN baseline (top half) and our DDPM (bottom half) model. Each row

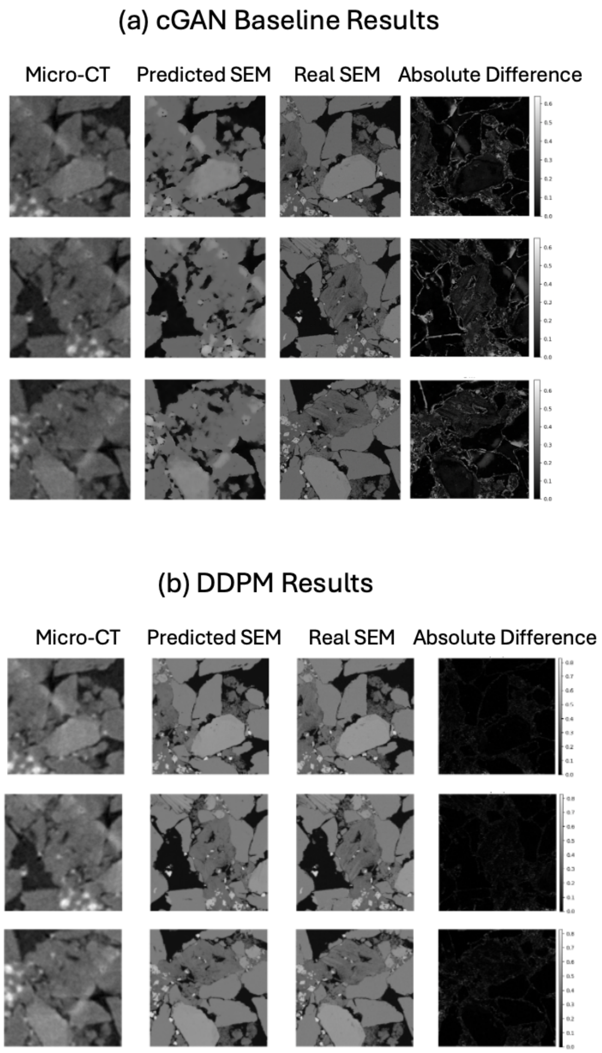

Figure 3. Comparison between the cGAN baseline (top half) and the DDPM (bottom half) on three randomly selected test samples.

comprises four panels: the leftmost column shows the low-resolution micro-CT input tile, the second column displays the corresponding predicted SEM, the third column is the ground-truth SEM, and the rightmost column shows the pixel-wise absolute-difference image $|\hat{x} - x|$ by gray scale.

In the cGAN results (Figure 3, top half), the generator recovers the general grain boundaries but fails to capture finer intragranular textures. In the first row, for example, sub-micron cracks visible in the true SEM appear noticeably attenuated in the predicted SEM, and this loss of detail is highlighted by the bright, irregular patterns in the absolute-difference map. The second row reveals that regions of dark matrix material are over-smoothed, causing pores to blur together; the corresponding difference image shows a dense texture of mid-tone errors around these blurred boundaries. In the third row, the cGAN again washes out high-frequency variations within individual mineral grains, producing larger, smoother patches where the true SEM exhibits subtle tonal shifts. Consequently, the absolute-difference map for row 3 is characterized by widespread speckled noise, indicating that many small-scale features were not faithfully reconstructed.

By contrast, the DDPM outputs (Figure 3, bottom half) demonstrate markedly sharper and more accurate recovery of fine-scale SEM textures. In the first row, cellular-scale crack networks and intragranular inclusions present in the ground truth are rendered with high fidelity, resulting in an almost uniformly dark absolute-difference image (indicating very low pixel-wise error). The second row further illustrates this point: pores and mineral inclusions maintain crisp edges, and the DDPM's difference map shows only faint, localized errors where minor shading variations occur. In the third row, heterogeneous grain-face textures—such as etched surfaces and small inclusions—match the true SEM almost exactly, and the difference panel is nearly blank except for a few isolated bright spots where residual noise remains.

Overall, these three test samples reveal that the DDPM more consistently preserves high-frequency SEM features and produces substantially lower reconstruction error than the cGAN. The cGAN's difference images exhibit pronounced textured patterns of error, whereas the DDPM's difference plots are almost uniformly dark, confirming that our diffusion-based approach captures sub-micron detail that the GAN baseline tends to miss.

5.4. Discussion

Hyperparameter Choices For the cGAN, we adopted a learning rate of 2×10^{-4} with Adam ($\beta_1 = 0.5, \beta_2 = 0.999$), which aligns with best practices in Pix2Pix-style image-to-image translation [7]. This value provided a stable adversarial training regime while maintaining reasonable convergence speed. A mini-batch size of 24 was selected to fully utilize the H100 GPU memory without sacrificing gradient diversity. For the DDPM, we reduced the learning rate to 5×10^{-5} and introduced a weight decay of 5×10^{-4} (via AdamW) to accommodate the longer diffusion chain and discourage over-smoothing during early denoising iterations. We also applied gradient clipping at 0.7 and used a ReduceLROnPlateau scheduler on validation loss. These hyperparameter settings were determined through grid-search-style tuning on the 4,800-pair validation set and yielded the lowest validation loss before plateauing (around epoch 150 for the cGAN and epoch 300 for the DDPM).

Failure Modes Although the DDPM substantially outperforms the cGAN baseline in both quantitative metrics and visual fidelity, it is not without failure cases. The most common limitations arise when attempting to reconstruct extremely fine or novel lithologies that the micro-CT input cannot resolve. In a few test samples, needle-shaped mineral flakes on the order of 100 nm still appear as low-contrast streaks in the predicted SEM (see Figure 3, bottom). This occurs because the micro-CT voxels (2 μm resolution) do not contain any signal for sub-100 nm features—hence, the diffusion model can only predict plausible texture based on the training distribution. By comparison, the cGAN frequently suffers from over-smoothing and missing pores in low-contrast regions, which indicates mode collapse of fine-scale textures. In short, while the DDPM recovers high-frequency detail more faithfully, it remains constrained by the fundamental resolution limits of the micro-CT modality.

Overfitting and Generalization Both models were trained exclusively on rotation-augmented tiles drawn from a single North Sea sandstone core and then tested on held-out tiles (19 rotations per original 512×512 patch). Consequently, the reported metrics and visual examples (e.g. Figures 2 and 3) reflect performance on the same geological specimen, although on never-seen rotations and spatial locations. To mitigate overfitting, we employed aggressive rotation-safe augmentation and early-stopping based on validation SSIM/PSNR. Both training and validation curves stabilized well before the final epochs (cGAN by epoch 150, DDPM by epoch 300), with no signs of divergence. Nonetheless, true cross-sample generalization to different rock types or mineralogies remains untested. Future work should validate these architectures on SEM/micro-CT pairs from other cores to ensure robustness across a broader range of pore-scale textures.

Summary In summary, our results demonstrate that a conditional DDPM dramatically outperforms a Pix2Pix cGAN baseline on the micro-CT $\rightarrow$ SEM super-resolution task. Quantitatively, the DDPM achieves mean SSIM 0.802 and PSNR 27.08 dB on 7,985 unseen test pairs—compared to 0.595 SSIM and 20.52 dB PSNR for the cGAN (Table 1). Qualitatively, the DDPM preserves sub-micron intragranular features and complex pore geometries that the cGAN tends to smooth out (Figure 3). Therefore DDPM's superior ability to predict realistic SEM textures from micro-CT input suggests it is well-suited for non-destructive, high-fidelity pore-scale characterization of geological materials, especially in scenarios where accurate SEM-level detail is required.

6. Conclusion and Future Work

In this work, we compared a Pix2Pix-style conditional GAN (cGAN) against a conditional diffusion model (DDPM) for super-resolving micro-CT images into SEM-quality reconstructions. Using a heavily augmented dataset drawn from a single North Sea sandstone slice, we demonstrated that the DDPM significantly outperforms the cGAN baseline on unseen test pairs. Image comparisons confirm that the diffusion model recovers fine pore-scale textures and mineral boundaries better, while the cGAN often produces oversmoothed outputs.

However, our dataset derives entirely from one rock sample, with diversity introduced through rotation-safe tiling rather than distinct lithologies. Consequently, future work should incorporate truly independent cores and varied rock types to assess generalization. Additionally, we can investigate semi-supervised or unsupervised approaches to reduce the reliance on perfectly paired micro-CT/SEM data, for example by incorporating cycle-consistency or contrastive learning losses that can leverage large unpaired collections of micro-CT images. Overall, these findings indicate that conditional diffusion models offer substantial potential for non-destructive, high-fidelity micro-CT $\rightarrow$ SEM super-resolution.

7. Acknowledgements

We gratefully acknowledge the creators of the Pix2Pix reference implementation (`pytorch-CycleGAN-and-pix2pix`) [2] and Lucidrains's `denoising-diffusion-pytorch` codebase [3], which formed the foundation of our cGAN and DDPM implementations, respectively. We also thank the authors of the Hugging Face "Annotated Diffusion" blog [4] for their clear exposition of diffusion model concepts and implementation details, which greatly aided our adaptation. Some of the computing for this project was performed on the Sherlock cluster. We would like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to these research results. Finally, we appreciate the support and feedback from our TA and classmates during the development and evaluation of this project.

References

[1] D. Brunet, E. R. Vrscay, and Z. Wang. On the mathematical properties of the structural similarity index. *IEEE Transactions on Image Processing*, 21(4):1488–1499, 2011.

[2] G. Buono, S. Caliro, G. Macedonio, V. Allocca, F. Gamba, and L. Pappalardo. Exploring microstructure and petrophysical properties of microporous volcanic rocks through 3d multiscale and super-resolution imaging. *Scientific Reports*, 13(1):6651, 2023.

[2] `https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix.git`

[3] `https://github.com/lucidrains/denoising-diffusion-pytorch.git`

[4] `https://huggingface.co/blog/annotated-diffusion`

[3] Y. Da Wang, R. T. Armstrong, and P. Mostaghimi. Boosting resolution and recovering texture of micro-ct images with deep learning. *arXiv preprint arXiv:1907.07131*, 2019.

[4] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 184–199. Springer, 2014.

[5] O. Furat, D. P. Finegan, Z. Yang, T. Kirstein, K. Smith, and V. Schmidt. Super-resolving microscopy images of li-ion electrodes for fine-feature quantification using generative adversarial networks. *npj Computational Materials*, 8(1):68, 2022.

[6] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[8] S. J. Jackson, Y. Niu, S. Manoorkar, P. Mostaghimi, and R. T. Armstrong. Deep learning of multi-resolution x-ray micro-ct images for multi-scale modelling. *arXiv preprint arXiv:2111.01270*, 2021.

[9] J. Korhonen and J. You. Peak signal-to-noise ratio revisited: Is simple beautiful? In *2012 Fourth international workshop on quality of multimedia experience*, pages 37–38. IEEE, 2012.

[10] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[11] S. C. Park, M. K. Park, and M. G. Kang. Super-resolution image reconstruction: a technical overview. *IEEE signal processing magazine*, 20(3):21–36, 2003.

[12] A. Roslin, M. Marsh, N. Piché, B. Provencher, T. Mitchell, I. Onederra, and C. Leonardi. Processing of micro-ct images of granodiorite rock samples using convolutional neural networks (cnn), part i: Super-resolution enhancement using a 3d cnn. *Minerals Engineering*, 188:107748, 2022.

[13] L. Shan, C. Liu, Y. Liu, Y. Tu, S. V. Chilukoti, and X. Hei. Single image multi-scale enhancement for rock micro-ct super-resolution using residual u-net. *Applied Computing and Geosciences*, 22:100165, 2024.

[14] E. Ugolkov, X. He, H. Kwak, and H. Hoteit. Memory-efficient super-resolution of 3d micro-ct images using octree-based gans: Enhancing resolution and segmentation accuracy. *arXiv preprint arXiv:2505.18664*, 2025.

[15] E. Ugolkov, X. He, H. Kwak, and H. Hoteit. Super-resolution of 3d micro-ct images using generative adversarial networks: Enhancing resolution and segmentation accuracy. *arXiv preprint arXiv:2501.06939*, 2025.

[16] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.