Real-World Super-Resolution for μCT Rock Imaging Using Simple Real-ESRGAN-3D

Yue Sun*

Department of earth science and engineering, Imperial College London, London, England, United Kingdom

Abstract. Computer tomography technology is widely used in geological exploration as a non-destructive 3D imaging modality integrated with computer simulation. However, high-resolution CT scanning is constrained by economic costs and long acquisition times, and higher resolution often reduces the effective field-of-view in practice. Computational super-resolution (SR) offers a scalable alternative by enhancing low-resolution scans and reducing the need for additional acquisitions. The paper presents Simple Real-ESRGAN-3D, a GAN-free, × 2 volumetric SR model for micro-CT (μ CT) data. The network is a lightweight residual 3D CNN with trilinear upsampling, trained end-to-end using content-only losses: L1, MSE, and a slice-consistency L1 term that encourages coherence across adjacent slices. This design improves training stability and lowers computational cost while preserving fine pore-scale details. On representative μ CT rock volumes, the method delivers strong SR quality, demonstrating efficient and reliable enhancement from low-resolution inputs. By upgrading image fidelity without adversarial training, Simple Real-ESRGAN-3D enables more efficient pore-scale structural analysis and numerical simulation from standard μ CT acquisitions.

1 Introduction

Using digital rock technology to study the physical properties of reservoir rocks, pore-throat network structure, and the micro-seepage mechanism of fluid in porous media provides a new perspective for the accurate characterization of rock pores. Digital rock physics (DRP) can use high - resolution tomographic images of rock samples to model and predict physical properties such as porosity, permeability, and elastic moduli [1]. In Digital rock physics, X-ray micro-CT is a powerful tool for reservoir analysis because it can non-destructively resolve the complex pore structure of rocks at the micron scale. In particular, high-resolution imaging captures pore networks and throat sizes that directly influence permeability and fluid transport in reservoir rocks.

However, practical CT imaging involves a fundamental trade-off between resolution, field-of-view, and scan time [2]. High-resolution imaging requires small sample volumes or long scan times, limiting statistical representativeness. Multiple scans at different

* Corresponding author: yue.sun25@imperial.ac.uk

magnifications are often needed to gain better fine pore details and each extra acquisition will add cost and delays. So computational SR methods are attractive because they can enhance resolution in silico without new imaging [3].

Super-resolution (SR) aims to reconstruct a plausible high-resolution (HR) image from a corresponding low-resolution (LR) observation. In the past decade, deep convolutional neural networks (CNNs) have transformed SR for natural images: early work such as SRCNN demonstrated that even a shallow CNN can learn effective upsampling with strong fidelity. Subsequent residual and GAN-based architectures further improved perceptual quality and recovery of fine detail.

Several recent works have specifically targeted digital rock SR. Wang et al. (2018) introduced a 3D-SRCNN for CT rock volumes that uses voxel-wise 3D convolutions with residual learning to perform multi-scale SR, and stabilized training with practical tactics—adjustable learning rates, gradient clipping, and momentum SGD—to address slow convergence and heavy GPU memory demands. Beyond supervised 3D CNNs, 2D CNN/GAN frameworks have been adapted to micro-CT rocks. ESRGAN employs residual-in-residual dense blocks to recover realistic textures, while Real-ESRGAN extends this idea by explicitly modelling complex degradations and improving stability with spectral normalization and a U-Net discriminator; its training is conducted on fully synthetic pairs generated by a high-order blur/downsample pipeline. In digital-rock settings, both ESRGAN and Real-ESRGAN have been reported to sharpen pore boundaries and enhance mineral/pore segmentation by restoring high-frequency details [4]. To reduce reliance on perfectly paired LR–HR volumes, unpaired/unsupervised GANs have also been explored. Niu et al. (2022) compared paired CNNs with an unpaired GAN (CinCGAN) on carbonate datasets and found that both approaches can reconstruct SR images that preserve petrophysical accuracy, with the unpaired GAN offering the practical advantage of training on arbitrary, unpaired collections.

Despite these advances, current SR approaches have limitations. Volumetric deep networks (3D CNNs or 3D GANs) contain millions of parameters and demand large GPU memory and long training times [5]. For instance, training a 3D CNN requires careful cropping of CT volumes and complex learning-rate schedules to fit hardware constraints. Similarly, GAN-based SR methods often require elaborate discriminator designs (e.g., U-Net with skip connections) and stabilization tricks (spectral normalization, progressive training) to converge. Such requirements lead to high computational cost and slow experimentation.

To address these challenges, the paper introduce Simple Real-ESRGAN-3D, a lightweight 3D super-resolution model tailored to reservoir μCT data. The network stacks residual blocks of Conv3D–GroupNorm–ReLU and deliberately omits any adversarial objectives [6]. The research targets a fixed ×2 upscaling of pore-space volumes (5 μm → 2.5 μm voxels), training solely on synthetic LR–HR pairs produced by applying a known Gaussian blur followed by downsampling. Optimization is purely content-driven—combining L1, MSE, and a slice-consistency L1 computed on random axial slices—to promote volumetric coherence. This design simplifies training, reduces model complexity, and explicitly enforces inter-slice consistency. Trained end-to-end on paired volumes without adversarial or perceptual losses, the model focuses on faithfully reconstructing pore-space intensities while minimizing artifacts.

The key contributions of this work include:

- A simple 3D CNN architecture for volumetric super-resolution of rock micro-CT images, avoiding the complexity of GANs while incorporating Group Normalization and residual connections for stable deep learning [7].

● A fixed 2× super-resolution framework for 3D μCT data (5 μm → 2.5 μm) based on a single, explicit degradation model (Gaussian blur + downsampling), enabling straightforward generation of LR–HR training pairs [8].

● A content-only loss strategy (L1 and MSE) augmented with a novel slice-consistency loss on random axial sections, designed to preserve inter-slice consistency in the super-resolved volume [9].

2 Dataset and Preprocessing

The paper evaluates on 3D μ CT porous rock data using the Bentheimer_1000c_3p0035um sandstone volume (Imperial College London; ∼1000^3 voxels, grayscale). The raw stack is loaded as a 3D array and min–max normalized to [0,1].

To build 2× SR pairs, the research synthesizes LR volumes from the HR data with a single isotropic degradation: apply Gaussian blur (σ =0.8voxels) to mimic optical blur, then decimate by 2 along each axis (simple stride-2 subsampling). The paper adds no extra noise or compression, so LR–HR differences mainly reflect resolution loss. This controlled setup keeps the pipeline simple, in contrast to Real-ESRGAN's multi-stage degradation. For efficiency, the experiment extracts paired patches offline: HR patches of 64^3 voxels and corresponding LR patches of 32^3 voxels, sampled from diverse locations while keeping LR/HR grids aligned (starts snapped to multiples of 2). All pairs are stored in a single HDF5 file with HR and LR datasets and gzip compression. During training, patches are loaded as tensors and normalized per tensor (consistent with the code), enabling fully supervised learning of the ×2 upscaling from LR inputs to HR targets under this single-degradation setting.

3 3D Super-Resolution Model Architecture

The research generator is a compact residual 3D CNN that maps a 32^(3) LR patch to a 64^3 SR patch. It comprises: 1.Stem: one 3×3×3 Conv3d (in_channels = 1, out_channels = 80). 2.Body: 12 residual blocks, each with Conv3d–GroupNorm–ReLU ×2 and an additive skip (no dense connections; no BatchNorm). 3.Feature fusion: a Conv3d with a global residual skip from the stem features.4. Pre-upsampling: a light module before scaling.5. Reconstruction head: final prediction layers.[10]

For x2 upscaling, the paper applies Conv3d + ReLU followed by trilinear interpolation (scale_factor = 2, align_corners = False) to avoid checkerboard artifacts. A terminal 3×3×3 Conv3d plus sigmoid yields a single-channel output in [0,1] [11]. The design is conceptually related to ESRGAN/Real-ESRGAN but is GAN-free—there is no discriminator and no adversarial loss [12]. This removes adversarial instability and heavy hyper-parameter tuning. Instead, the experiment trains with content-only losses (L1, MSE, and a slice-consistency L1 on random axial slices), which preserves fine detail and delivers strong perceptual quality [13].

4 Training Loss Functions

The research trains the generator without any adversarial loss, using a weighted sum of three content losses. Let I_{SR} denote the super-resolved output and I_{SR} the ground-truth volume (both normalized to [0,1]).

Mean Squared Error (MSE):

$$L_{MSE} = \| I_{SR} - I_{HR} \|_2^2 \quad (1)$$

MSE correlates with PSNR and penalizes larger deviations [14].

Mean Absolute Error (L1):

$$L_1 = \| I_{SR} - I_{HR} \|_1 \quad (2)$$

L1 is less sensitive to outliers and helps preserve sharper edges than MSE alone.

Slice L1 (axial).

To encourage slice-wise consistency, the paper computes an L1 loss on random axial (z) slices. Given a depth D, if D > 4 sample four distinct indices$S = \{i_1, \dots, i_4\} \subset \{0, \dots, D-1\}$ and define

$$L_{slice} = \frac{1}{|S|} \sum_{i \in S} \| I_{SR}[:,:,i,:,:] - I_{HR}[:,:,i,:,:] \|_1 \quad (3)$$

If $D \leq 4$, it will fall back to full-volume L1. This term gently regularizes the network to produce coherent details when viewed slice-by-slice along depth.

The total loss is

$$L_{total} = L_1 + 0.1 L_{MSE} + 0.02 L_{slice} \quad (4)$$

with fixed, empirically chosen weights: L1 carries the most weight to preserve pixel fidelity, MSE serves as a lighter PSNR-oriented term, and a small slice-consistency penalty acts as a stability-friendly perceptual proxy. Training is strictly GAN-free. The experiment is optimized with AdamW (learning rate 3×10^{-4}, weight decay 1×10^{-4}), use a cosine annealing schedule, and enable mixed precision (AMP) for efficiency and stability. For evaluation, the result reports SSIM and LPIPS; no adversarial loss or VGG-based perceptual loss is used during training.

5 Baseline (Non-learning)

The experiment uses trilinear interpolation as the 2× baseline. Each LR patch is upsampled by trilinear interpolation to the HR size, and metrics are computed against the corresponding HR target. This matches the interpolation primitive used within the model's upsampling path. All results are reported on held-out μCT sub-volumes that were never seen during training.

6 Metrics and Qualitative Assessment

The paper evaluates fidelity and perceptual quality with PSNR, SSIM, and LPIPS, and complement them with visual inspection of 3D renderings and 2D slices. PSNR (dB) measures distortion via the mean-squared error and is reported on volumes normalized to [0,1]; higher is better, though it can favor overly smooth results. SSIM quantifies structural similarity (luminance, contrast, structure) and better reflects preservation of pore–throat geometry; the paper computes it on normalized volumes (or orthogonal slices) and average over held-out sub-volumes—again, higher is better. LPIPS is a learned perceptual distance that correlates with human judgments of detail; the paper computes it slice-wise and averages across slices and volumes, where lower is better. For qualitative comparison, it examines opacity-weighted 3D volume renderings and representative axial/coronal/sagittal slices,

focusing on pore boundary sharpness, continuity of thin features, aliasing/staircasing, and potential hallucinations.

7 Dataset and Preprocessing

On the held-out μ CT sub-volumes, the model outperforms the non-learning baseline (trilinear interpolation) on all metrics. PSNR rises from 18.06 dB to 20.12 dB (+2.06 dB), SSIM increases from 0.926 to 0.943 (+0.017), and LPIPS (AlexNet) drops from 0.288 to 0.213 (−0.075, ~26% lower). These gains indicate sharper pore boundaries, better structural consistency, and reduced perceptual error relative to simple interpolation. In Qualitative observations, the metric gains align with clear visual improvements grain pore interfaces appear crisper, thin throats remain continuous (fewer breaks or merges), and small cavities emerge without amplifying noise. The +2.06 dB PSNR reflects lower overall error, the SSIM +0.017 tracks better preservation of pore geometry, and the LPIPS −0.075 matches the perceived boost in edge sharpness and texture realism—achieved without hallucinated details thanks to the GAN-free training [15].

8 Results and Discussion

The super-resolution approach yields a notable improvement in reconstruction quality, with PSNR increasing from 18.06 dB to 20.12 dB, SSIM rising from 0.926 to 0.943, and LPIPS (AlexNet) dropping from 0.288 to 0.213. This ~2 dB gain in PSNR indicates significantly lower pixel-wise error, while the higher SSIM and lower LPIPS reflect better preservation of structural details and perceptual similarity to the ground truth. In other words, the model improves both fidelity and perceptual quality relative to the baseline. Despite these gains, several limitations remain, which point to avenues for further improvement:

8.1 Evaluation pipeline:

Small inconsistencies can distort PSNR. The results observed that half-voxel misalignment between the synthetic LR generation (Gaussian blur + ×2 decimation) and the upsampling grid, or improper boundary handling, can depress volumetric PSNR. Even a half-pixel offset or failing to crop blurred borders introduces a subtle global shift that inflates MSE and thus lowers PSNR. Strict grid alignment, consistent normalization, and explicit border cropping are therefore essential (many benchmarks crop output borders for this reason; e.g., the ESA Kelvins challenge).

The research will tighten downsampling/upsampling alignment and standardize normalization and border handling so that PSNR reflects true reconstruction quality rather than evaluation artifacts.

8.2 Loss function balance:

The model deliberately weighted the loss toward L1, with a smaller MSE term and a light slice-wise L1. This favors edge and texture preservation, which matches the strong SSIM/LPIPS gains observed. The trade-off is expected: because PSNR is tightly linked to MSE, giving L1 most of the weight leaves some global error on the table compared with pure MSE minimization. Optimizing only MSE typically raises PSNR but often over-smooths fine details.

The research plans to explore a more balanced or adaptive scheme—e.g., begin with L1-heavy training to learn structure, then increase the MSE weight (or switch to a light two-stage

schedule) to drive down residual error. The goal is to lift PSNR further without sacrificing the perceptual and structural quality captured by SSIM/LPIPS.

8.3 Capacity and context limitations:

The model trains on small 3D patches (input 32^3 → output64^3) with a single-scale architecture, which limits the receptive field and the geological context visible at once. As a result, heterogeneous textures and large structures—such as pore networks or mineral grains extending beyond 64 voxels—may be only partially captured. In practice, this can cause minor seams at patch boundaries and makes it harder to super-resolve features larger than the patch.

The research will expand the model's effective context by (i) adopting multi-scale or hierarchical SR (e.g., MS-ResUNet-style fusion across scales), (ii) training with larger patches, and (iii) using overlap-tile inference to blend boundaries. These steps increase the receptive field and help the network learn long-range correlations, enabling more consistent reconstruction of large or spatially complex structures.

8.4 Simplified degradation model:

The model trained with a single, controlled degradation—fixed Gaussian blur and noise—to synthesize LR inputs. This makes training stable but does not cover the variety of point-spread functions, noise spectra, or artifacts found in real μCT/microscopy data. When test images deviate from these assumptions, a domain gap arises: PSNR (being MSE-based) drops sharply with even small intensity or blur mismatches, while SSIM and LPIPS—which emphasize structure and perception—are more tolerant to moderate shifts.

To reduce this gap, the research will adopt a more diverse and robust degradation pipeline: mix multiple blur kernels and noise levels, or use blind SR that learns degradations directly. The research will also explore domain adaptation, e.g., fine-tuning on a small set of real LR/HR pairs (when available) or using unsupervised adaptation to real data. These steps should improve real-case PSNR and overall reliability in unseen conditions.

In summary, addressing these limitations forms the core of the future work. The research will improve the evaluation procedure and loss function balance to squeeze out higher fidelity, enhance the network architecture and training strategy to capture larger context, make the model more robust to real-world degradations, and incorporate domain-specific validations. These future improvements are expected to further boost the accuracy and practical utility of super-resolution for rock images, ultimately enabling more reliable digital rock analyses and predictions

9 Conclusion

The research presented Simple Real-ESRGAN-3D, a lightweight 2× super-resolution method for 3D μCT rock imaging that emphasizes simplicity and robustness. The model uses a compact residual 3D CNN with trilinear upsampling and is trained with only content losses (L1, weighted MSE, and slice-consistency L1) under a classical Gaussian blur + downsampling degradation model. By avoiding any GAN or adversarial loss, the approach yields stable, interpretable results; content-driven design remains reliable. On held-out test volumes, the paper observes consistent quantitative gains: PSNR improves from 18.06 to 20.12 dB, SSIM from 0.926 to 0.943, and LPIPS drops from 0.288 to 0.213 (a 26.1 % reduction). These improvements in standard image quality metrics, which are commonly

used in CNN-based SR evaluation, demonstrate that the simple model matches or exceeds paired CNN baselines (which typically yield higher PSNR/SSIM than unpaired GAN models)

Crucially, the design trades adversarial complexity for reproducibility and stability – a valuable property in digital rock science workflows. Recent work shows that 3D CNN SR outputs can produce physically meaningful results (e.g. porosity and permeability consistent with high-resolution scans) when properly validated. Going forward, the research plans to enrich the degradation model (to better mimic real scanner blur and noise) and experiment with perceptual/feature-space losses. The research will also pursue task-driven validation, assessing how SR images improve downstream analysis such as pore segmentation, topology characterization, and flow permeability estimates. All these extensions will further bolster the utility of research simple, robust SR framework for 3D rock imaging.

References

1. M.J. Blunt, B. Bijeljic, H. Dong, O. Gharbi, S. Iglauer, P. Mostaghimi, A. Paluszny, C. Pentland, Pore-scale imaging and modelling. Adv. Water Resour. **51**, 197–216 (2013)
2. X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, C.C. Loy, ESRGAN: Enhanced super-resolution generative adversarial networks, in Proceedings of the European Conference on Computer Vision (ECCV) Workshops (2018)
3. X. Wang, L. Xie, C. Dong, Y. Shan, Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data, in Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (AIM@ICCV) (2021)
4. Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
5. R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 586–595 (2018)
6. Micro-CT Images and Networks (Bentheimer sandstone: 1000^3, 3.0035 μm), Imperial College London, 2015. Retrieved from https://www.imperial.ac.uk/earth-science/research/research-groups/pore-scale-modelling/micro-ct-images-and-networks/(2025)
7. A. Rabbani, M. Babaei, R. Shams, Y.D. Wang, T. Chung, DeePore: A deep learning workflow for rapid and comprehensive characterization of porous materials. Adv. Water Resour. **146**, 103787 (2020)
8. J. de C.V. Fernandes, V.A.V. Vidal, A.D. Vidal, R.M. Barros, Absolute permeability estimation from microtomography rock images through deep learning super-resolution and adversarial fine tuning. Sci. Rep. **14**, 16704 (2024)
9. Y. Wang, Q. Teng, X. He, J. Feng, CT-image of rock samples super resolution using 3D convolutional neural network. Comput. Geosci. **133**, 104314 (2019)
10. A. Roslin, S. Baskoro, R. Kusumaningrum, Processing of micro-CT images of granodiorite rock samples using a 3D convolutional neural network super-resolution algorithm. Comput. Geosci. (2023)
11. M. Bizhani, M.H. Bodaghi, H. Hosseini-Nasab, M.A. Javadi, Reconstructing high fidelity digital rock images using deep learning. Sci. Rep. **12**, 4489 (2022)
12. G. Buono, S. Caliro, G. Macedonio, V. Allocca, F. Gamba, L. Pappalardo, Exploring microstructure and petrophysical properties of microporous volcanic rocks through 3D multiscale and super-resolution imaging. Sci. Rep. **13**, 6651 (2023)

13. M. Li, S. Shen, W. Gao, W. Hsu, J. Cong, Computed Tomography Image Enhancement using 3D Convolutional Neural Network, in DLMIA/ML-CDS Workshops (MICCAI) (2018)
14. B. Lim, S. Son, H. Kim, S. Nah, K.M. Lee, Enhanced Deep Residual Networks for Single Image Super-Resolution, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (NTIRE), 136–144 (2017)
15. C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, … W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4681–4690 (2017)