Mathematical Geosciences manuscript No.
(will be inserted by the editor)

HLF-SASR: A High–Low Frequency Guided Structure-Aware Super-Resolution Network for Digital Rock Images

Tianyu Gao[1] · **Chenyang Zhu**[1] (✉) · **Zhenwei Niu**[1] · **Mingjie Li**[1] · **Yunxin Xie**[1] · **Fang Wang**[2] ·

Received: date / Accepted: date

Abstract High-resolution digital rock images provide the foundation for pore-scale reservoir characterization, seepage simulation, and multi-scale flow analysis. However, the inherent trade-off between field of view and spatial resolution in micro-computed tomography limits the acquisition of high-resolution images over macro-scale rock samples. Deep learning-based super-resolution offers a way to mitigate this limitation by learning mappings from low-resolution to high-resolution images, aiming to recover fine-scale pore–throat structures while preserving a large field of view. Yet existing super-resolution methods still struggle with the strong structural heterogeneity and frequency characteristics of rock images, often producing blurred pore boundaries or over-smoothed mineral textures. This paper proposes a High–Low Frequency guided Structure-Aware Super-Resolution network (HLF-SASR), which combines a component-aware backbone with a stacked hourglass backbone with an ASPP-based multi-scale context module, a high–low frequency collaborative attention module, and a temperature-scaled multi-expert fusion strategy. Experiments on a public digital rock dataset using two-dimensional sandstone and carbonate slices, where 200×200 and 400×400 image pairs are used as low-resolution and high-resolution inputs for $2\times$ super-resolution, show that HLF-SASR consistently surpasses representative super-resolution models such as Enhanced Deep Super-Resolution network (EDSR), Residual Channel Attention Network (RCAN), and Multi Attention Super-Resolution Neural Network (MASR) in terms of peak signal-to-noise ratio and structural similarity index. Visual comparisons further demonstrate that the proposed network produces clearer and more continuous pore–throat boundaries, better preserves narrow fractures and sharp-corner structures, and avoids over-sharpening mineral matrices or introducing pseudo-textures, thereby providing more reliable reconstructions for digital rock analysis.

[1] School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou, 213000, China
[2] Department of Computer Science, Brunel University London, London, UB8 3PH, United Kingdom
E-mail: zcy@cczu.edu.cn

Keywords Digital core · Image SR · Multi-scale context · High-low frequency attention · Temperature-scaled Softmax fusion

1 Introduction

Digital core technology acquires three-dimensional pore structures from Micro-Computed Tomography (micro-CT) and couples them with numerical flow simulations, thereby supporting reservoir evaluation, pore-scale mechanism analysis, and multi-scale flow studies. However, micro-CT imaging is fundamentally constrained by a trade-off between Field Of View (FOV) and spatial resolution: a larger FOV typically implies lower spatial resolution, whereas High-Resolution (HR) scans are difficult to apply to macro-scale rock samples. As a result, it is challenging to simultaneously capture large-scale structural integrity and fine pore–throat details within a single scan, which substantially limits the potential of digital core analysis for reservoir characterization and rock physics studies (Cnudde and Boone 2013).

Deep learning-based super-resolution (SR) has recently emerged as a promising means to mitigate this limitation, as deep visual models have demonstrated strong capability across a wide range of image analysis and structural understanding tasks (Otoom and Etoom 2025).By learning the mapping from Low-Resolution (LR) to HR images, SR models can, in principle, recover micron-scale pore–throat structures while maintaining a large FOV, thus improving the geometric fidelity of digital cores (Dong et al. 2016). Although convolutional neural networks and residual SR architectures have achieved remarkable success on natural images, direct application to microscopic rock images remains suboptimal. The mismatch is mainly reflected in the following three aspects.

First, rock images exhibit a highly heterogeneous frequency distribution. Structures such as pore throats and fractures correspond to high-frequency components, whereas mineral matrices and background textures are dominated by low-frequency content. If all frequency components are processed together in a mixed feature space, high-frequency structures are easily overwhelmed by low-frequency responses, leading to blurred pore–throat boundaries and attenuated textures. Most existing SR models lack an explicit high-/low-frequency disentangling mechanism and therefore struggle to adapt to the complex frequency-domain characteristics of digital rock images.

Second, pore systems in natural lithologies span multiple spatial scales, ranging from micron-scale micropores to millimeter-scale dissolution pores within the same sample. Single-scale convolutions have difficulty jointly capturing long-range connectivity and local texture details, which has become a key bottleneck for further improving SR performance (Chen et al. 2017 Zhao et al. 2017). Although multi-scale feature extraction has been extensively validated in semantic segmentation and high-level vision, systematic integration of multi-scale context modeling into 2D SR networks for digital cores remains limited.

Third, from a microstructural viewpoint, different regions in micro-CT and thin-section images exhibit substantially different textures and mechanical behaviors. Flat mineral substrates, pore boundary zones, and fractures or pore throats near sharp corners present distinct morphologies and stress concentration characteristics (Liu et al.

2025). In deep visual models, mainstream channel–spatial attention and feature fusion modules often rely on Sigmoid-normalized scalars to weight branches or channels. Such mechanisms use fixed activation forms and lack explicit temperature or sparsity control (Hu et al. 2018). When lithological differences are large or noise levels vary, these fixed gating mechanisms are difficult to tune for fine-grained control of the relative contributions of different feature branches, which limits the adaptability of the model to complex pore structures.

To address these challenges, this work proposes a High–Low Frequency Spatial Attention Super-Resolution network (HLF-SASR) tailored for digital core SR reconstruction, incorporating three key improvements:

1. A high/low-frequency attention mechanism that explicitly splits feature channels into low-frequency and high-frequency subspaces and processes mineral backgrounds and pore–throat structures in separate branches, thereby enhancing discrimination between structural details and background regions.
2. A multi-scale context fusion module that introduces Atrous Spatial Pyramid Pooling (ASPP) at the bottleneck of the network. Convolutions with multiple dilation rates are used to capture structural dependencies across scales, effectively improving the simultaneous recovery of both large pores and fine micro-pore throats.
3. A temperature-scaled Softmax fusion strategy that normalizes the outputs of multiple expert branches, enabling dynamically adjustable structural selection and collaboration and enhancing the generalization capability of the model across different lithologies and noise conditions.

The remainder of this paper is organized as follows. Section 2 reviews related work on deep learning-based SR for digital rock images. Section 3 presents the HLF-SASR framework, including the overall architecture and core modules. Section 4 reports experimental results, compares the proposed method with mainstream SR approaches, and analyzes the contribution of each module through ablation studies. Finally, Section 5 concludes the work and discusses future research directions.

2 Related Work

2.1 Overall Methods for Rock Image Reconstruction

Paired SR methods rely on accurately registered LR-HR samples for supervised learning. Niu et al. (Niu et al. 2021) compared the performance of paired convolutional neural networks and unpaired generative adversarial networks on carbonate rock micro-CT datasets, demonstrating that GAN-based unpaired SR is a feasible alternative when paired data are unavailable, albeit with slightly inferior physical consistency indices compared with paired approaches. More recent work has predominantly adopted real paired micro-CT or SEM data for training. For example, the Multi Attention Super-Resolution Neural Network (MASR) framework proposed by Xing et al. (2023) showed that paired supervision markedly improves the recovery of pore–throat details and connectivity.

Unpaired SR approaches typically employ generative adversarial networks to learn the mapping between LR and HR domains without requiring one-to-one correspondence. For instance, SRCycleGAN reconstructs HR images from LR inputs using cycle-consistency constraints (Chen et al. 2020). Such methods are suitable when paired data are extremely scarce; however, their training is often unstable and it is difficult to guarantee strict physical connectivity and consistency of the reconstructed pore network.

Deep residual networks constitute the mainstream architecture for single-image SR. Enhanced Deep Super-Resolution network (EDSR) (Lim et al. 2017) achieves state-of-the-art performance on natural image SR by removing batch normalization layers and increasing the depth of residual blocks, while Residual Channel Attention Network (RCAN) (Zhang et al. 2018) introduces channel attention within a residual-in-residual framework and significantly enhances the recovery of high-frequency details. Nevertheless, when directly applied to rock images, these networks face challenges caused by heterogeneous mineral textures, strong imaging noise, and multi-scale pore systems. Their generic design, optimized for natural images, often fails to fully capture rock-specific structural patterns and pore connectivity.

GAN-based SR further aims to generate perceptually realistic textures via adversarial and perceptual losses. Super-Resolution Generative Adversarial Network (SRGAN) (Ledig et al. 2017) and Enhanced Super-Resolution Generative Adversarial Networks (ESRGAN) (Wang et al. 2019a) achieve high perceptual quality on natural images, but adversarial training can introduce non-physical artifacts and hallucinated structures. Such artifacts are particularly problematic for digital rock applications, where pore–throat connectivity and geometric fidelity are critical for subsequent flow simulations. To improve physical consistency, several recent studies have incorporated additional constraints such as gradient loss, Hessian loss, and power spectrum regularization into the SR objective (Zhu et al. 2023 Li et al. 2024). These efforts underscore the importance of combining data-driven reconstruction with structure-aware and physics-related priors for reliable rock image SR, consistent with the broader trend of deep learning applications in diverse scientific imaging and perception settings (Zhu et al. 2025 Xiao and Dong 2025).

2.2 High-Low Frequency Attention Mechanism

Rock thin-section and micro-CT images typically contain both high-frequency structures, such as pore-throat boundaries and fractures, and low-frequency regions dominated by mineral substrates. When a network processes all frequency components in a unified manner, high-frequency details are easily suppressed by low-frequency responses, leading to edge over-smoothing and loss of fine textures. To address this issue, an increasing number of studies have explored explicit high-low frequency decomposition and attention modeling in various vision tasks (Ha et al. 2024 Ren et al. 2025).

For example, Huang et al. (2017) proposed Wavelet-SRNet, which employs wavelet decomposition to partition images into multiple frequency bands and models each band separately, thereby significantly improving the recovery of high-frequency de-

tails. Guo et al. (2023) introduced the Spatial-Frequency Attention Network (SFANet), which first obtains spectral representations via two-dimensional Fourier transform and then adaptively reweights different frequency bands through a frequency-channel attention module to finely control frequency responses. In a related but distinct setting, Zhang et al. (2025a) proposed the LOFI framework for facial expression recognition with noisy labels, where attention is modeled as a dynamic process evolving with training iterations and combined with spatial structural information to gradually rearrange attention distributions, thus mitigating the impact of label noise.

In scientific imaging and image SR, several works have explicitly emphasized high-frequency structures or enforced frequency-domain consistency at the loss-function level. Chen et al. (2026), for instance, introduced a gradient-based edge loss for dental cone-beam CT SR, using HR micro-CT as a reference to highlight fine anatomical structures such as tooth roots and trabecular bone. Fuoli et al. (2021) directly measured spectral discrepancies between reconstructed images and ground truth in the frequency domain and improved the perceptual quality of textures and high-frequency content through a Fourier-space loss.

However, SR studies dedicated to digital cores remain relatively limited in terms of frequency-domain modeling. Most existing methods still process features uniformly in the spatial domain and do not explicitly exploit the distinct behaviors of high- and low-frequency structures. As a consequence, pore-throat boundaries are often under-reconstructed, and local textures tend to be excessively smoothed, which restricts the fidelity of pore-scale characterization in digital rock applications.

2.3 Multi-Scale Structure and Temperature-Controlled Fusion

Pore structures and mineral grain sizes in rock images span a wide range of spatial scales. Single-scale convolutions struggle to simultaneously capture the long-range connectivity of pore networks and the fine-grained textures of local micropores. Consequently, multi-scale feature extraction has become a key strategy for improving reconstruction quality. In high-level vision, DeepLab employs ASPP to model multi-scale context via parallel atrous convolutions with different dilation rates, achieving outstanding performance in semantic segmentation (Chen et al. 2018), while PSP-Net aggregates global information across scales using pyramid pooling (Zhao et al. 2017). For image restoration, Multi-Scale Residual Network (MSRN) (Li et al. 2018) embeds multi-scale convolution kernels within residual blocks and effectively enhances feature expressiveness. However, these multi-scale designs are primarily tailored to natural images or semantic-level tasks, and their context modeling capacity remains limited when applied to rock images with intricate pore networks. For instance, although U-Net preserves information at multiple scales through skip connections, it still faces difficulties in jointly capturing long-range semantic context and high-frequency details such as thin pore throats and sharp corners.

Several recent studies have introduced multi-scale architectures into reconstruction and analysis pipelines for digital core images. Shan et al. (2024) proposed a multi-scale enhanced residual U-Net for single-image SR of rock micro-CT images, which improves both the clarity and connectivity of pore–throat boundaries. Zhang

et al. (2025b) further developed a lightweight multi-scale attention feature distillation network to extract and fuse texture and pore structures at different scales in sandstone and carbonate rocks, achieving high-fidelity SR on digital core CT data. In segmentation tasks, Wang et al. (2022) adopted the UNet++ architecture with dense multi-scale skip connections, significantly improving boundary localization and the detection of small pore targets. Despite these advances, systematic integration of ASPP-like atrous convolution spatial pyramid pooling into 2D rock SR backbones remains limited. This gap leaves room for further exploration of how to jointly leverage long-range dependencies and local textures in digital core SR.

Regarding feature fusion, selective integration of multi-branch representations has been shown to improve model adaptability and robustness. In classical Mixture-of-Experts (MoE) models, Softmax gating is used to interpret attention weights as a probability distribution over experts and to generate a data-dependent weighted combination of their outputs (Jordan and Jacobs 1994). The Gumbel-Softmax trick proposed by Jang et al. (2016) introduces a temperature parameter to control the sharpness of the distribution, enabling approximately discrete selection of one or a few experts while retaining differentiability. In image restoration, the Adaptive Sparse Transformer proposed by Zhou et al. (2024) performs adaptive routing and sparse feature selection along spatial and channel dimensions, suppressing redundant responses and emphasizing salient structures. These studies collectively demonstrate that learnable Softmax-based gating, MoE-style fusion, and multi-branch feature routing can achieve data-adaptive selection among multiple pathways.

Overall, existing work still leaves significant room for improvement in both multi-scale modeling and dynamic fusion for digital rock SR. The application of multi-scale context modules within 2D SR backbones is not yet fully exploited, and temperature-scaled Softmax fusion strategies have not been systematically investigated in component-aware digital core SR frameworks.

3 Methodology

3.1 Network Architecture Overview

The proposed HLF-SASR network adopts a stacked hourglass backbone to perform multi-scale feature extraction and iterative refinement for rock image super-resolution. The hourglass module follows a symmetric encoder–decoder design with successive downsampling and upsampling stages (Xing et al. 2023). Along the bottom-up path, spatial resolution is progressively reduced to aggregate coarse, global contextual information; along the top-down path, features are upsampled and fused across scales, typically via skip connections that bridge the encoder and decoder. Multiple hourglass modules are stacked sequentially so that each stage refines the prediction of the previous stage, which has been shown to improve reconstruction accuracy (Wei et al. 2020). In HLF-SASR, the final stage outputs the super-resolved rock image, while intermediate stages provide auxiliary predictions for deep supervision to facilitate stable optimization and progressive refinement.

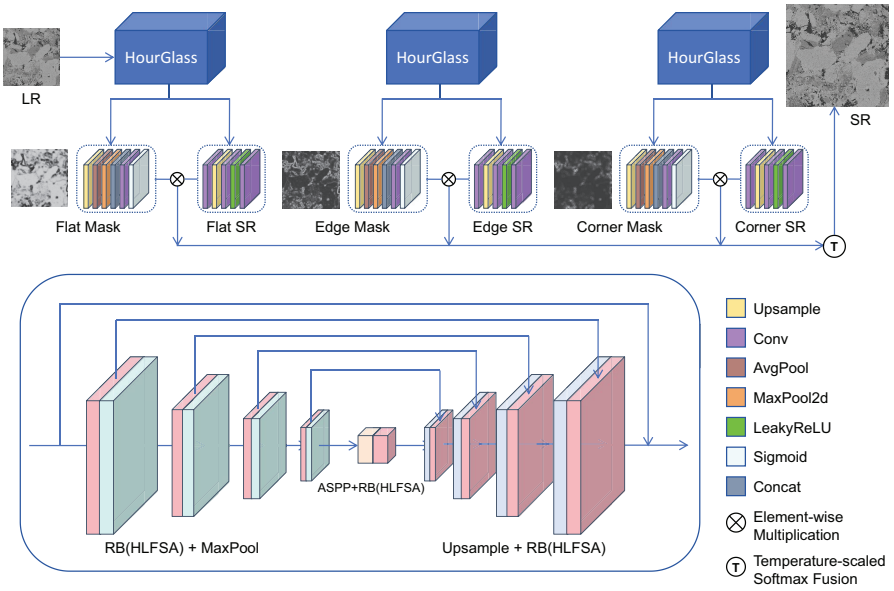

Fig. 1: Overview of the HLF-SASR framework. A stacked HourGlass backbone produces component-aware SR outputs whose predictions are fused by temperature-scaled Softmax, while residual blocks with HLFSA and an ASPP bottleneck capture frequency-aware, multi-scale features.

On top of this backbone, several task-specific modules are introduced to adapt the architecture to the characteristics of rock image SR. First, a High–Low Frequency Spatial Attention (HLFSA) module is incorporated into the residual blocks to modulate feature responses in a frequency-aware manner, enhancing discriminative details in structurally complex regions while suppressing noise in homogeneous areas. Second, an ASPP bottleneck is employed to capture multi-scale contextual cues with different receptive fields, further enriching the representation of heterogeneous rock structures. Finally, a multi-expert fusion strategy, governed by a temperature-scaled Softmax gating function, assigns different structural patterns to specialized expert branches and combines their outputs into a unified SR prediction. As illustrated in Fig. 1, the stacked hourglass backbone produces component-aware SR outputs, which are adaptively fused by the learned gating weights to yield the final high-quality reconstruction.

For convenience, the main symbols specific to the proposed HLF-SASR framework are summarized in Table 1.

3.2 High–Low Frequency Spatial Attention Module

To enhance the network's ability to focus on structurally critical regions, a HLFSA mechanism is introduced that explicitly fuses low-frequency global context with high-

Table 1: Summary of main notations in the Methodology section.

Symbol	Description
$\mathbf{F}$	Intermediate feature map in the network
$\mathbf{F}^L, \mathbf{F}^H$	Low- and high-frequency sub-features obtained by channel split
$\mathbf{G}$	Low-frequency response map in the HLFSA module
$\mathbf{H}$	High-frequency response map in the HLFSA module
$\mathbf{A}$	Single-channel spatial attention map produced by HLFSA module
$\mathbf{F}'$	Attention-refined feature
$\mathbf{F}_{\text{ASPP}}$	Bottleneck feature enhanced by the ASPP module
$\mathbf{F}_{\text{bot}}$	Bottleneck feature map at the lowest spatial resolution in the SR backbone
$\mathbf{S}$	Gating score tensor for the three experts: flat, edge and corner
$W_k(n,h,w)$	Normalized Softmax gating weight of the k-th expert at (n,h,w)
T	Temperature parameter controlling the sharpness of Softmax gating
N	Batch size (number of input images)
C	Number of feature channels
H	Height of the feature map
W	Width of the feature map

frequency local details. This design is motivated by the observation that low-frequency components primarily encode global structure and intensity distribution, whereas high-frequency components capture edges, corners, and fine textures. Given an intermediate feature map $\mathbf{F} \in \mathbb{R}^{N\times C\times H\times W}$, HLFSA processes $\mathbf{F}$ through two complementary branches-one modeling global structure and the other emphasizing local detail-and then fuses their outputs to obtain a refined spatial attention map. By explicitly separating and adaptively combining frequency-related cues, the module produces a more discriminative attention distribution that guides precise feature reweighting.

Formally, the module takes $\mathbf{F} \in \mathbb{R}^{N\times C\times H\times W}$ as input and outputs an attention map $\mathbf{A} \in \mathbb{R}^{N\times 1\times H\times W}$ that only varies over spatial locations. The attention map is broadcast along the channel dimension and used to uniformly weight all channels at each spatial position, leading to an enhanced feature $\mathbf{F}'$ with the same size as $\mathbf{F}$. To capture frequency-aware behavior, the design exploits the fact that different channels tend to learn responses corresponding to different frequency bands and explicitly splits $\mathbf{F}$ along the channel dimension into two sub-features: $\mathbf{F}^L \in \mathbb{R}^{N\times C_L\times H\times W}$ is interpreted as a low-frequency subspace that mainly describes large-scale mineral distributions and slowly varying backgrounds, while $\mathbf{F}^H \in \mathbb{R}^{N\times C_H\times H\times W}$ is interpreted as a high-frequency subspace focusing on rapidly changing structures such as pores, throats, and fractures. The channel split is defined as $C_L$and C_H. where $C_L$is given by Eq. (1),

$$C_L = \left\lfloor \frac{C}{2} \right\rfloor, \tag{1}$$

and C_H is given by Eq. (2),

$$C_H = C - C_L, \tag{2}$$

where C denotes the total number of channels of the input feature map $\mathbf{F}$.

This simple bisection avoids the overhead of explicit frequency-domain transforms while reserving independent modeling capacity for the subsequent low- and

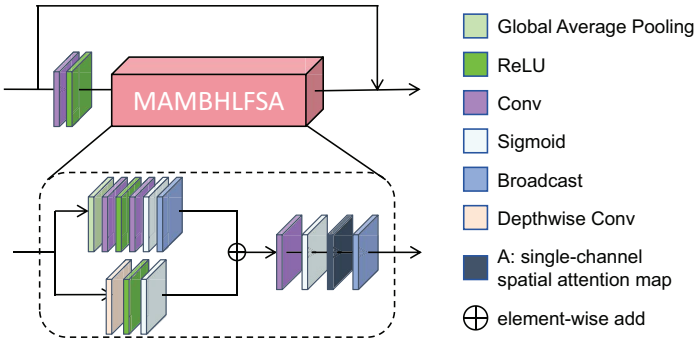

Fig. 2: Architecture of the proposed HLFSA, which combines global-average low-frequency context and depthwise high-frequency.

high-frequency branches, enabling them to specialize in complementary structural patterns.

The low-frequency branch operates on $\mathbf{F}^L$, which mainly encodes large-scale, slowly varying information such as mineral backgrounds and bedding structures. To aggregate global context, global average pooling (GAP) is first applied to compress the spatial information of each channel into a single statistic, and the resulting global descriptors are then passed through a lightweight Multi-Layer Perceptron (MLP) implemented as two successive 1×1 convolutions with non-linear activation, yielding low-frequency attention weights as Eq. (3),

$$\mathbf{g} = \sigma\big(W_2\,\delta\big(W_1\,\mathrm{GAP}(\mathbf{F}^L)\big)\big), \tag{3}$$

where $\delta(\cdot)$ denotes the ReLU activation, $\sigma(\cdot)$ is the Sigmoid function, and W_1, W_2 are the learnable parameters of the 1×1 convolutions. The resulting $\mathbf{g} \in \mathbb{R}^{N\times C_L\times 1\times 1}$ is then broadcast along the spatial dimensions to form $\mathbf{G} \in \mathbb{R}^{N\times C_L\times H\times W}$, which serves as a low-frequency response map with spatially smooth attention patterns within each channel. This map emphasizes globally important regions while preserving structural coherence across the rock sample.

The high-frequency branch focuses on local, rapidly changing structures in $\mathbf{F}^H$. To extract such details, a 3×3 depthwise convolution is applied, followed by batch normalization, ReLU activation, and a Sigmoid function, as shown in Eq. (4),

$$\mathbf{H} = \sigma\big(\delta\big(\mathrm{BN}(\mathrm{DWConv}_{3\times 3}(\mathbf{F}^H))\big)\big), \tag{4}$$

where $\mathbf{H} \in \mathbb{R}^{N\times C_H\times H\times W}$, $\mathrm{DWConv}_{3\times 3}(\cdot)$ denotes a depthwise convolution applied independently to each channel, $\mathrm{BN}(\cdot)$ is batch normalization. This chain of operations enables each high-frequency channel to act as a learnable high-pass filter: batch normalization and non-linear activation stabilize and enhance the representational capacity, while the Sigmoid compresses responses into $[0,1]$, making the high-frequency attention more robust and easily compatible with the low-frequency branch. As a result, $\mathbf{H}$ highlights fine-scale structures such as edges, corners, and pore-throat boundaries.

To obtain a unified spatial attention map, the channel dimensions of the two branches are first aligned using a 1×1 convolution applied to the high-frequency response. Specifically, $\mathbf{H}$ is projected to the same number of channels as $\mathbf{G}$ by a 1×1 convolution with kernel W_h, element-wise addition with $\mathbf{G}$ is then performed, and the fused feature is finally compressed to a single-channel map through another 1×1 convolution followed by Sigmoid activation, as shown in Eq. (5),

$$\mathbf{A} = \sigma\big(W_m * (\mathbf{G} + W_h * \mathbf{H})\big), \tag{5}$$

where W_h and W_m denote the 1×1 convolution kernels applied to the high-frequency response and the fused feature, respectively, $*$ denotes convolution, and $\mathbf{A}$ is the final single-channel spatial attention map. In this way, $\mathbf{A}$ jointly encodes the smooth global structural information from the low-frequency branch and the localized detail emphasis from the high-frequency branch, providing a unified and interpretable attention distribution for subsequent spatial weighting.

Finally, the spatial attention map $\mathbf{A}$ is broadcast along the channel dimension and used to modulate the original feature map through element-wise multiplication. In other words, the refined feature map $\mathbf{F}'$ is obtained by treating $\mathbf{A}$ as a spatial weight mask and applying the element-wise product. Since $\mathbf{A}$ only depends on spatial positions, this operation assigns a unified importance weight to each location (h, w) across all channels. In high-frequency salient regions such as pores, throats, and fractures, larger values of $\mathbf{A}(h, w)$ amplify local responses and facilitate the reconstruction of fine structural details. Conversely, in homogeneous mineral substrates or background regions, smaller values of $\mathbf{A}(h, w)$ suppress noise and pseudo-textures, thereby preserving the smoothness and stability of low-frequency structures.

3.3 Multi-scale Context Enhancement

While the stacked hourglass backbone excels at capturing features across different resolution levels through its symmetric encoder-decoder structure, its multi-scale capability is primarily hierarchical and sequential. At the deepest bottleneck layer, each position in the feature map $\mathbf{F}_{\text{bot}}$ still responds within a relatively limited and fixed receptive field. This limits the network‘s ability to integrate broad spatial contextual information at a single semantic level, which is crucial for understanding long-range interactions within complex pore-throat systems.

To complement the hourglass's hierarchical multi-scale representation with parallel multi-context aggregation, we insert an ASPP module at the bottleneck. ASPP operates on the same feature map $\mathbf{F}_{\text{bot}}$ but employs parallel atrous convolutions to explicitly gather contextual information from multiple spatial scales simultaneously. This provides a rich structural prior that is orthogonal to the hourglass‘s design, specifically enhancing the modeling of long-range dependencies before upsampling.

In single-image SR, the network must simultaneously capture fine-grained texture details and large-scale structural context within a limited FOV. This requirement is particularly pronounced in digital rock imaging, where pore-throat systems exhibit strong multi-scale characteristics and span a wide range of spatial extents. Relying solely on stacked standard convolutions at the bottleneck layer makes it difficult to

obtain a sufficiently large effective receptive field, which limits the ability to model long-range dependencies in the LR space. To alleviate this limitation, an ASPP module is inserted at the hourglass bottleneck to aggregate multi-scale contextual information from the encoded features, providing a structural prior for subsequent upsampling and reconstruction.

Let the bottleneck feature be denoted by $\mathbf{F}_{\text{bot}} \in \mathbb{R}^{N\times C\times H\times W}$. ASPP constructs multiple parallel atrous convolution branches with different dilation rates $r_k k = 1^K$ on the same feature map. For a depthwise 3×3 atrous convolution with dilation rate r, its discrete form can be written as Eq. (6),

$$y_c(p) = \sum_{\Delta\in\Omega} w_c(\Delta), F_c\big(p + r\cdot\Delta\big), \tag{6}$$

where p denotes a spatial position, Ω is the sampling set of the 3×3 convolution kernel, $w_c(\Delta)$ is the kernel weight at offset Δ for the c-th channel, and $F_c(\cdot)$ denotes the c-th input channel.

Consequently, ASPP enables multi-scale modeling, where small dilation rates r focus on very local textures and large dilation rates capture large-scale structures, which is beneficial for encoding long-range interactions between pores, throats, and mineral matrices already in the LR feature space.

In the proposed implementation, four dilation branches with $r \in 1,2,4,8$ are employed, and a lightweight depthwise separable ASPP design is adopted to control computational complexity. For each dilation rate r_k, the branch consists of a depthwise 3×3 convolution with dilation r_k followed by a pointwise 1×1 convolution and a LeakyReLU activation, as expressed in Eq. (7),

$$U_{r_k} = \phi\Big(\text{PWConv}_{r_k}^{1\times 1}\big(\text{DWConv}_{r_k}^{3\times 3}(\mathbf{F}_{\text{bot}})\big)\Big), \tag{7}$$

$\text{DWConv}^{3\times 3} r_k$ denotes a 3×3 depthwise convolution with dilation rate r_k, $\text{PWConv}^{1\times 1} r_k$ denotes a 1×1 pointwise convolution, $\phi(\cdot)$ is the LeakyReLU activation function, and $U r_k$ is the output of the k-th dilation branch. Here $\text{DWConv}_{r_k}^{3\times 3}$ extends the 3×3 depthwise convolution used in the high-frequency branch by introducing a dilation rate r_k.

The resulting multi-scale features $U_{r_k} k = 1^4$ are first aggregated by element-wise summation and then linearly fused by a 1×1 convolution $\text{Conv}_f^{1\times 1}$. The fused feature is added back to the original bottleneck feature in a residual manner to form the ASPP-enhanced representation, as shown in Eq. (8),

$$\mathbf{F}_{\text{ASPP}} = \mathbf{F}_{\text{bot}} + \text{Conv}_f^{1\times 1}\left(\sum_{k=1}^{4} U_{r_k}\right). \tag{8}$$

Within the hourglass-based backbone, $\mathbf{F}_{\text{ASPP}}$ is propagated from the bottleneck to the subsequent top-down decoding and upsampling branches as an enhanced latent feature. Compared with simply stacking standard convolutions, the ASPP module explicitly expands the effective receptive field at negligible additional parameter and computational cost, enabling the SR reconstruction process to simultaneously exploit local details such as edges and corners and broader structural constraints across

heterogeneous pore systems. This multi-scale context enhancement is particularly advantageous for accurately restoring complex rock microstructures in digital core SR.

3.4 Temperature-scaled Softmax Gating Fusion Strategy

In a typical component-aware SR branch, three experts produce structure-specific SR predictions, denoted by $\mathrm{SR}^{\mathrm{flat}}$, $\mathrm{SR}^{\mathrm{edge}}$, and $\mathrm{SR}^{\mathrm{corner}}$. These predictions are originally combined by non-exclusively weighting and summing them using three Sigmoid masks generated by CompMask. Although this design encourages specialization for different structure types, it exhibits two main limitations. First, multiple experts may yield high activations at the same spatial location, leading to substantial overlap in their effective regions and making it difficult to establish a clear spatial division of labor. Second, the masks are generated solely via independent Sigmoid activations, without an explicit mechanism to control the strength of competition among experts, which restricts the network's ability to adaptively modulate the relative dominance of each expert across spatial positions according to reconstruction demands.

To enhance structural discrimination during three-expert fusion, a temperature-scaled Softmax gating mechanism is introduced to replace the Sigmoid-based non-exclusive weighting. By enforcing a normalized distribution over experts at each spatial position and modulating its sharpness through a temperature parameter, the gating mechanism adaptively adjusts between competitive and collaborative expert behaviors. This leads to a clearer division of functional roles among experts and yields a more selective fusion scheme, which is particularly beneficial for reconstructing heterogeneous lithologies and complex pore structures in digital rock images.

In implementation, three independent gating branches are designed for the three structural types, each generating a gating score map with the same spatial resolution as the SR feature maps. Specifically, S^{flat} exhibits responses biased toward flat regions, S^{edge} toward edge regions, and S^{corner} toward corner or high-curvature structures. These three maps are stacked along the channel dimension to form $\mathbf{S} \in \mathbb{R}^{N\times 3\times H\times W}$, where the three channels correspond to the three experts and the spatial dimensions are aligned with the SR predictions. Before normalization, the channels of $\mathbf{S}$ can be interpreted as raw response intensities of flat, edge, and corner structures at each spatial location, indicating which structural pattern each position is most consistent with.

A temperature-scaled Softmax is then applied to $\mathbf{S}$ along the expert dimension to obtain per-pixel expert weights, as defined in Eq. (9),

$$W_k(n,h,w) = \frac{\exp\big(S_k(n,h,w)/T\big)}{\sum_{j=1}^{3}\exp\big(S_m(n,h,w)/T\big)}, \tag{9}$$

where n indexes images in the batch, and (h,w) denotes the spatial location, $S_k(n,h,w)$ is the gating score of the k-th expert at position (n,h,w), $T>0$ is the temperature parameter, and $W_k(n,h,w)$ is the normalized gating weight, satisfying $\sum_{k=1}^{3} W_k(n,h,w) =$

1 and $0 < W_k(n,h,w) < 1$. The fused SR output is obtained as a per-pixel mixture of expert predictions, as shown in Eq. (10),

$$I^{\mathrm{SR}}(n,c,h,w) = \sum_{k=1}^{3} W_k(n,h,w), I_k^{\mathrm{SR}}(n,c,h,w), \tag{10}$$

where c indexes channels, $k \in \{\mathrm{flat}, \mathrm{edge}, \mathrm{corner}\}$ and I_k^{SR} denotes the SR prediction of the k-th expert.

The temperature T plays a central role in controlling the selectivity of the fusion. When $T \to 0^+$, the Softmax distribution becomes highly peaked, and the weight at each spatial position is concentrated on a single expert, effectively yielding a hard assignment and enforcing a strong division of labor. When $T = 1$, the Softmax produces a moderately selective distribution that balances discrimination and smoothness. When $T > 1$, the distribution becomes flatter and the three experts receive more similar weights, promoting cooperative behavior among experts within the same spatial region.

By tuning T, the gating mechanism provides a continuous transition between expert specialization and collaborative fusion. This flexibility allows the model to adapt expert usage to varying lithologies, structural complexities, and noise levels, thereby improving the robustness and interpretability of component-aware SR in digital rock image reconstruction.

3.5 Training Pipeline

To enhance the fidelity of high-frequency structures such as edges and fine details in the reconstructed rock images, a hybrid loss is employed that combines a gradient-weighted reconstruction term with component-level supervision.

As a basic reference, the pixel-wise ℓ_1 loss between the super-resolved result I^{SR} and the HR ground truth I^{HR} is defined as Eq. (11),

$$\mathscr{L}_{\mathrm{pixel}} = \frac{1}{HWC} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{k=1}^{C} \left| I^{\mathrm{SR}}(i,j,k) - I^{\mathrm{HR}}(i,j,k) \right|, \tag{11}$$

where H, W, and C denote the height, width, and number of channels of the image, respectively.

To explicitly emphasize high-frequency regions, classic one-dimensional Sobel derivative kernels are used to approximate the horizontal and vertical gradients of an image $\mathbf{I}$, namely $K_x = [-1,,0,,1]$ and $K_y = [-1,,0,,1]^\top$. Applying these operators to I^{SR} and I^{HR} yields the gradient maps $\mathbf{G}x^{\mathrm{SR}}$, $\mathbf{G}y^{\mathrm{SR}}$, $\mathbf{G}x^{\mathrm{HR}}$, and $\mathbf{G}y^{\mathrm{HR}}$. Based on the gradient discrepancy at each spatial location (i,j), a spatial weight map is constructed as Eq. (12),

$$D_{\mathrm{GW}}(i,j) = \left(1 + \alpha \left| G_x^{\mathrm{SR}}(i,j) - G_x^{\mathrm{HR}}(i,j) \right|\right) + \left(1 + \alpha \left| G_y^{\mathrm{SR}}(i,j) - G_y^{\mathrm{HR}}(i,j) \right|\right), \tag{12}$$

where α is a hyperparameter controlling the emphasis on gradient discrepancy. The corresponding gradient-weighted reconstruction loss is defined as Eq. (13),

$$\mathscr{L}_{\mathrm{GW}} = \frac{1}{HWC} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{k=1}^{C} D_{\mathrm{GW}}(i,j) \left| I^{\mathrm{SR}}(i,j,k) - I^{\mathrm{HR}}(i,j,k) \right|, \tag{13}$$

where $D_{\mathrm{GW}}(i,j)$ is broadcast along the channel dimension. This loss assigns larger penalties to pixels whose gradients deviate more from the ground truth, thereby encouraging sharper and more accurate reconstruction of edges and other high-frequency details.

In addition to the global constraint, the network explicitly outputs super-resolved results I_i^{SR} for three structural components. Using precomputed masks M_i^{HR} derived from the HR image, component-level supervision is imposed as Eq. (14),

$$\mathscr{L}_i = \mathscr{L}_{\mathrm{pixel}} \left(I_i^{\mathrm{SR}} \odot M_i^{\mathrm{HR}}, I^{\mathrm{HR}} \odot M_i^{\mathrm{HR}} \right), \tag{14}$$

where $i \in \{\mathrm{flat}, \mathrm{edge}, \mathrm{corner}\}$ and $\odot$ denotes element-wise multiplication. This term enforces that each expert branch concentrates on accurately reconstructing its corresponding structural component.

It is important to emphasize that the flat, edge, and corner masks introduced here provide a soft structural prior for training, rather than pixel-wise accurate hard labels. These masks are computed from local image statistics, which inherently exhibit smoothing effects and are therefore robust to high-frequency noise. More importantly, the objective of the component-level loss is not to enforce strict classification within each masked region, but to impose a qualitative and trend-level constraint: encouraging smoothness in flat regions and promoting sharpness and clarity in edge and corner regions. As long as the masks can approximately capture the spatial distribution of different structural types, the supervision remains effective and robust. Minor inaccuracies near mask boundaries or noise-induced responses act as mild regularization and do not destabilize training or hinder convergence. Furthermore, these masks are only used during training to shape reconstruction preferences and are not involved in inference, and thus do not affect the model's generalization performance.

The final training objective is a weighted combination of the global gradient-weighted loss and the component-level losses, as shown in Eq. (15),

$$\mathscr{L}\mathrm{total} = \mathscr{L}_{\mathrm{GW}} + \sum_{i \in \{\mathrm{flat},\mathrm{edge},\mathrm{corner}\}} \lambda_i \mathscr{L}_i, \tag{15}$$

where λ_i are hyperparameters that balance the contribution of each structural component. This joint objective constrains overall image quality through gradient-aware global supervision while simultaneously imposing differentiated local supervision on distinct structural components, thereby improving the reconstruction of high-frequency details such as pore–throat edges and corners without sacrificing global fidelity.

4 Experiment

4.1 Experiment Settings

The experimental design comprises the dataset specification, model training configuration, and evaluation protocol. All experiments are conducted on the publicly

available Digital Rock Super Resolution Dataset 1 (DRSRD1) released by Wang et al. (Wang et al. 2019b). DRSRD1 is constructed from micro-CT scans of Bentheimer sandstone and Estaillades carbonate and provides systematically organized 2D slices and 3D volumes. For each sample, co-registered HR images of size 800×800 and their $2\times$ and $4\times$ downsampled counterparts under a unified FOV are available, which makes the dataset suitable for SR tasks with different magnification factors.

In this study, the focus is on the 2D SR reconstruction setting. Two-dimensional sandstone and carbonate slices are selected from DRSRD1, and only the resolution pair 200×200 and 400×400 is used as the LR-HR input-target pair for the $2\times$ SR task. The original 800×800 images serve solely as HR background references for qualitative visualization and are not involved in training or validation. The paired LR-HR images are randomly split into training and test sets to ensure statistically independent evaluation. During training, standard data augmentation operations, including random cropping and horizontal/vertical flipping, are applied. Pixel intensities are normalized to the $[0,1]$ interval so that samples from different lithologies are compared on a consistent numerical scale.

In all experiments, 200×200 LR images from DRSRD1 are fed to the network, and the corresponding 400×400 HR images serve as supervision targets. The proposed HLF-SASR model adopts an hourglass-based backbone with 6 stacked HourGlass modules and 20 residual blocks, and is trained end-to-end for 250 epochs with a batch size of 16 using a fixed LR patch size of 48. Optimization is performed using Adam with an initial learning rate of 2×10^{-4} and $(\beta_1, \beta_2) = (0.9, 0.999)$. The learning rate is decayed using a piecewise step schedule, being multiplied by 0.5 at epochs 160, 200, and 230. To prevent gradient explosion, gradient clipping with a maximum ℓ_2 norm of 20 is applied to the generator parameters. During training, PSNR and the reconstruction loss are computed on a held-out validation set every 3 epochs, and the checkpoint with the highest validation PSNR is selected as the final model for testing.

Our HLF-SASR model is implemented as a stacked hourglass backbone with $n_{\mathrm{HG}} = 6$ hourglass modules. Each hourglass follows a symmetric encoder–decoder design with four downsampling stages using MaxPool with a scale factor of $\times 2$, followed by four corresponding upsampling stages implemented via bilinear interpolation with the same $\times 2$ scaling. The feature width along the encoder increases as $64 \to 96 \to 128 \to 160$ and remains 160 at the bottleneck. An ASPP module is inserted at the bottleneck with dilation rates $\{1, 2, 4, 8\}$.

The proposed high–low frequency spatial attention (HLFSA) is embedded into the residual units of both encoder and decoder blocks. HLFSA splits channels into low- and high-frequency subspaces, applies GAP and MLP to the low-frequency branch and depthwise 3×3 convolution to the high-frequency branch, and produces a single-channel spatial attention map shared across channels. We use reduction ratio $r = 4$, kernel size 3 and dilation 1 in all experiments.

For component-aware reconstruction, we employ three expert SR heads tapped from the 1st, 2nd and the last hourglass modules. Each expert produces an SR prediction using bilinear upsampling and convolutional refinement, and generates a corresponding confidence mask. The three masks are stacked and normalized with a temperature-scaled Softmax to fuse the expert outputs. A lightweight refinement head is optionally applied with $\alpha = 0.1$.

4.2 Quantitative Comparison with State-of-the-Art Methods

For quantitative evaluation, the proposed HLF-SASR model is compared with three representative SR baselines: EDSR (Lim et al. 2017), RCAN (Zhang et al. 2018), and MASR (Xing et al. 2023). All methods are trained and tested under the same data partition and training protocol on the carbonate and sandstone subsets of DRSRD1, performing $2\times$ SR reconstruction from 200×200 to 400×400 resolution. The average PSNR, SSIM, Porosity Error and Boundary Error are reported on the test set, and the results are summarized in Table 2.

Table 2: Quantitative comparison of different SR models on two rock types (carbonate and sandstone). PSNR: Peak Signal-to-Noise Ratio; SSIM: Structural Similarity Index Measure; Por. Err.: Porosity Error; Bnd. Err.: Boundary Fraction Error.

Sample	Model	**PSNR (dB)**	**SSIM**	**Por. Err. (%)**	**Bnd. Err. (%)**
Carbonate	RCAN Zhang et al. (2018)	27.5240	0.7488	1.790	8.735
	EDSR Lim et al. (2017)	27.2253	0.7302	1.811	8.716
	MASR Xing et al. (2023)	**28.7142**	0.7987	0.821	5.655
	HLF-SASR (Ours)	28.5488	**0.8041**	**0.416**	**4.540**
Sandstone	RCAN Zhang et al. (2018)	29.8416	0.8681	1.126	1.110
	EDSR Lim et al. (2017)	32.2041	0.8963	0.632	0.584
	MASR Xing et al. (2023)	33.2522	**0.9212**	0.131	0.454S
	HLF-SASR (Ours)	**34.2624**	0.9201	**0.114**	**0.241**

The quantitative results in Table 2 show that, for both carbonate and sandstone, the generic natural-image SR networks EDSR and RCAN are consistently outperformed by the digital-rock-oriented MASR and the proposed HLF-SASR across both pixel-wise fidelity and rock-physics indicators. For carbonate samples, RCAN and EDSR achieve PSNRs of 27.52 dB and 27.23 dB, with SSIM values of 0.7488 and 0.7302, respectively. However, their porosity errors remain around 1.8% and their boundary fraction errors are close to 8.7% , indicating noticeable distortion of pore-space morphology. MASR, which explicitly incorporates digital-rock priors, increases the PSNR to 28.71 dB and the SSIM to 0.7987, while nearly halving the porosity error to 0.821% and reducing the boundary fraction error to 5.655% . Building on MASR, HLF-SASR attains a comparable PSNR of 28.55 dB but further improves structural fidelity, yielding the highest SSIM and further suppressing porosity and boundary fraction errors to 0.416% and 4.540% , respectively. These results indicate more faithful pore-scale reconstruction and more accurate delineation of pore–matrix interfaces in carbonate rocks.

For sandstone samples, the performance gap between generic SR models and digital-rock-specific methods becomes even more pronounced. RCAN and EDSR obtain PSNRs of 29.84 dB and 32.20 dB with SSIMs of 0.8681 and 0.8963, respectively, yet still exhibit porosity errors in the range of 0.63–1.13% and boundary fraction errors between 0.58% and 1.11%. MASR substantially improves PSNR to 33.25 dB and achieves the highest SSIM of 0.9212, while simultaneously reducing the poros-

Table 3: Comparison of model complexity in terms of parameter count for different SR networks trained for 250 epochs.

Model	Number of parameters (Millions)	Training Epochs
EDSR Lim et al. (2017)	40.73	250
RCAN Zhang et al. (2018)	253.71	250
MASR Xing et al. (2023)	28.51	250
HLF-SASR (Ours)	**19.81**	250

ity and boundary fraction errors to 0.131% and 0.454%, establishing a strong digital-rock baseline. HLF-SASR further delivers the best overall performance by pushing PSNR to 34.26 dB and maintaining a comparable SSIM of 0.9201, while further decreasing the porosity and boundary fraction errors to 0.114% and 0.241%, respectively. These results indicate that, although MASR attains a slightly higher SSIM, the proposed HLF-SASR offers a more favorable trade-off between pixel-wise fidelity and rock-physics consistency, providing more reliable sandstone reconstructions for downstream digital core analysis.

As reported in Table 3, the proposed HLF-SASR model achieves the most favorable trade-off between reconstruction quality and model complexity. Specifically, HLF-SASR contains only 19.81M parameters, which is substantially fewer than MASR, with 28.51M parameters. It is also dramatically more compact than the generic architectures EDSR and RCAN, which contain 40.73M and 253.71M parameters, respectively. Given that all models are trained for the same number of epochs , these results demonstrate that HLF-SASR attains its performance gains with a significantly lower parameter budget, highlighting the effectiveness and parameter efficiency of the proposed frequency-aware and multi-scale design.

4.3 Ablation Study

This subsection investigates the contribution of the temperature-scaled Softmax gating, the HLFSA module, and the ASPP-based context module by gradually integrating them into a component-aware SR backbone, which serves as the baseline.In this baseline, the three expert branches are fused using independent Sigmoid masks generated by the CompMask module, without normalization across experts.

For the ablation study, intermediate supervision is enabled for all variants to stabilize the training of the multi-stage architecture. To improve computational efficiency, the training patch size is set to 60 and the batch size is increased to 32. As summarized in Table 4, all three modules yield clear and consistent improvements over the component-aware backbone on the DRSRD1 carbonate subset. The baseline achieves 26.74 dB PSNR and 0.7905 SSIM, while exhibiting relatively large porosity and boundary fraction errors of 2.49% and 2.84%, respectively, indicating noticeable distortion of the pore-space morphology.

Introducing only the temperature-scaled Softmax gating already brings a substantial gain. PSNR increases to 27.54 dB and SSIM to 0.8037, while the porosity error is reduced to 0.92% and the boundary fraction error to approximately 2.02%. This

Table 4: Ablation study of the temperature-scaled Softmax gating, HLFSA, and ASPP context module on the DRSRD1 carbonate dataset.

Variant	PSNR (dB)	SSIM	Por. Err. (%)	Bnd. Err. (%)
Component-aware backbone	26.7378	0.7905	2.4917	2.8375
Backbone + Softmax gating	27.5402	0.8037	0.9207	2.0153
Backbone + HLFSA	27.5846	**0.8060**	0.7757	**1.7907**
Backbone + ASPP context module	27.5854	0.8058	0.8718	1.8942
Full (gating + HLFSA + ASPP)	**27.5874**	0.8047	**0.7186**	2.0183

suggests that MoE style soft routing helps the network adaptively combine different reconstruction behaviors and better respect pore–matrix transitions.

When the HLFSA module is added to the backbone, the PSNR further rises to 27.58 dB and the SSIM reaches the highest value among all variants. At the same time, the porosity and boundary fraction errors drop markedly to 0.78% and 1.79%, respectively. These results indicate that explicit high–low frequency collaboration is particularly effective for preserving fine pore–throat details and sharp interfaces, which are crucial for accurate pore-scale characterization.

The ASPP-based context module produces a very similar PSNR and SSIM profile and also reduces the porosity and boundary fraction errors to 0.87% and 1.89%, respectively. This confirms that multi-scale context aggregation strengthens the global structural consistency of the reconstruction by integrating information over a broader receptive field.

In the full HLF-SASR configuration, where Softmax gating, HLFSA, and ASPP are jointly enabled, the PSNR is further pushed to 27.59 dB, the highest among all settings, and the porosity error is reduced to 0.72%, the lowest in the table. Although the SSIM and boundary fraction error about 2.02% are slightly less optimal than the best single-module variant, they still represent clear improvements over the backbone. Overall, these results demonstrate that the three modules are largely complementary: each individual component contributes a meaningful gain, and their combination achieves a favorable trade-off between pixel-wise fidelity and rock-physics consistency.

To further investigate the influence of key architectural hyperparameters, we conduct additional ablation experiments from channel ratio allocation across branches, the dilation rate parameter T, and different multi-scale dilation set combinations. The corresponding results are summarized in Tables 5, 6, and 7.

We first evaluate the impact of channel distribution among parallel branches. As shown in Table 5, the balanced allocation achieves the best overall performance, yielding the highest PSNR and SSIM as well as the lowest porosity prediction error. This indicates that an even channel distribution across receptive fields facilitates more effective structural representation.

Next, we analyze the effect of the single dilation parameter T. Table 6 shows that $T = 1$ provides the optimal reconstruction accuracy, achieving the highest PSNR and the lowest porosity error. Both smaller and larger dilation values degrade perfor-

Table 5: Ablation study on channel ratio allocation.

Channel Ratio	PSNR	SSIM	Por. Err. (%)	Bnd. Err. (%)
25:75	27.5330	0.8033	1.1331	2.0529
50:50	**27.5874**	**0.8047**	**0.7186**	2.0183
75:25	27.5001	0.8021	1.0683	**1.9848**

mance, suggesting that an appropriately sized receptive field is essential for effective feature aggregation.

Table 6: Ablation study on dilation rate parameter T.

T	PSNR	SSIM	Por. Err. (%)	Bnd. Err. (%)
0.1	27.2825	**0.8049**	0.9162	2.0785
0.5	27.4475	0.8007	1.5843	2.0100
1	**27.5874**	0.8047	**0.7186**	2.0183
2	27.4098	0.8031	1.2313	2.0420
5	27.4183	0.8024	1.0462	**1.8276**

Finally, we compare different multi-scale dilation combinations in the ASPP module. As shown in Table 7, the dilation set $(1,2,4,8)$ achieves the strongest comprehensive performance, yielding the highest PSNR and the lowest porosity and boundary-related errors. Although $(1,3,6,9)$ attains a marginally higher SSIM, the overall results suggest that exponentially spaced dilation rates provide more robust multi-scale feature representation.

Table 7: Ablation study on multi-scale dilation combinations.

Dilation Set	PSNR	SSIM	Por. Err. (%)	Bnd. Err. (%)
(1,2,3,4)	27.3392	0.8064	1.1576	2.0361
(1,3,6,9)	27.2567	**0.8066**	1.1545	2.0459
(1,4,8,12)	27.4219	0.8023	1.1235	2.0354
(1,2,4,8)	**27.5874**	0.8047	**0.7186**	**2.0183**

In summary, these ablation studies confirm that balanced channel allocation, an appropriate dilation parameter, and carefully designed multi-scale receptive field configurations are all crucial for improving reconstruction accuracy and enhancing the reliability of microstructural parameter estimation in rock image super-resolution.

4.4 Qualitative Analysis

The visualization results for carbonate samples are presented in Fig. 3. The original FOV and the location of the zoomed-in region are indicated in the upper-left sub-

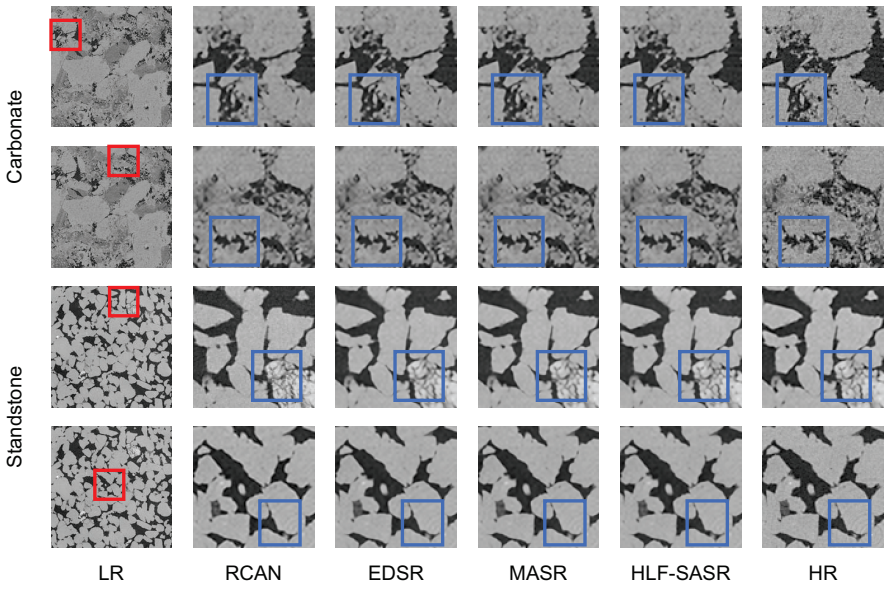

Fig. 3: Qualitative comparison of $\times 2$ SR results on carbonate samples using different SR models.

figure, while the remaining subfigures provide a side-by-side comparison of the reconstructed local region obtained by different SR models. RCAN and EDSR exhibit pronounced over-smoothing at the interfaces between complex pores and mineral matrices: pore-throat boundaries are blurred into smooth gray-level transition zones, and some small pores are partially or completely filled. MASR alleviates these artifacts and improves the recovery of pore contours; however, slightly rounded pore edges and local sticking of fine textures can still be observed. In contrast, the reconstruction produced by HLF-SASR is more consistent with the HR reference image: pore-throat boundaries remain clear and continuous, sharp corners and narrow pores are well preserved, and the interiors of mineral grains remain relatively smooth without noticeable noise amplification. As a result, HLF-SASR simultaneously avoids the over-smoothing observed in RCAN and EDSR and reduces local pseudo-textures compared with MASR. These qualitative observations indicate that, after introducing HLFSA and ASPP-based multi-scale context modeling, the proposed method provides superior boundary sharpness and detail integrity for complex pore structures in carbonate rocks.

To improve interpretability, we further visualize the spatial attention maps produced by the HLFSA module. As illustrated in Fig. 4, the attention responses are clearly concentrated around pore spaces, throat connections, and fracture boundaries rather than uniformly distributed over the rock matrix. In both carbonate and sandstone samples, the high-response regions align well with the segmented pore struc-

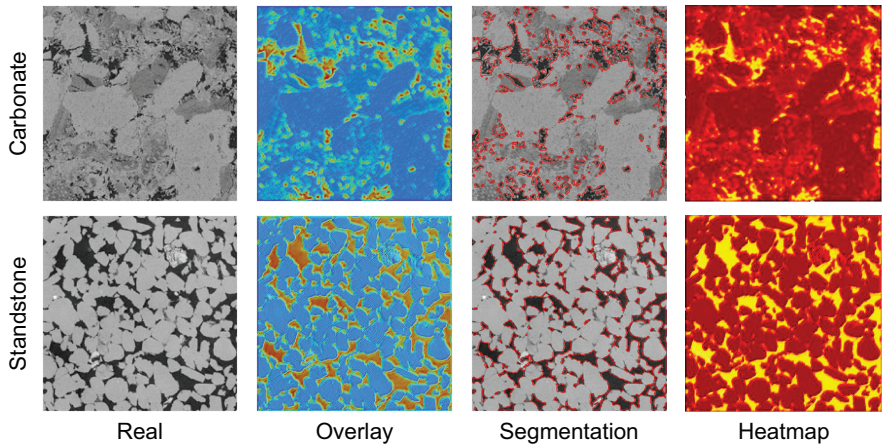

Fig. 4: Visualization of the spatial attention maps generated by the HLFSA module. From left to right: original image, attention overlay, segmented pore structure, and attention heatmap. Higher responses are clearly concentrated around pore spaces, throats, and fracture boundaries for both carbonate and sandstone samples.

tures, indicating that HLFSA effectively enhances the representation of flow-related heterogeneities and suppresses less informative background regions.

5 Conclusions

This study presents a frequency-aware and multi-scale enhanced SR network, termed HLF-SASR, for digital rock image reconstruction. The architecture integrates a HLFSA module, an ASPP-based multi-scale context enhancement module, and a temperature-scaled Softmax gating strategy into a component-aware SR backbone. By explicitly separating and collaboratively processing high- and low-frequency information, and by adaptvely routing structural components to specialized experts, HLF-SASR effectively improves the reconstruction of fine structures such as pore-throat boundaries and fractures while preserving global lithological context.

Extensive experiments on sandstone and carbonate samples from DRSRD1 demonstrate that HLF-SASR consistently outperforms representative state-of-the-art SR models in terms of PSNR and SSIM. Ablation studies further verify the effectiveness of each proposed module: temperature-scaled Softmax gating enhances the structural selectivity of multi-expert fusion, HLFSA strengthens frequency-aware feature modulation, and the ASPP context module enlarges the effective receptive field to better capture multi-scale pore systems. Qualitative analyses show that the proposed method yields clearer pore-throat boundaries, better preserves narrow fractures and sharp-corner structures, and suppresses pseudo-textures, leading to visually more faithful and geologically plausible reconstructions.

Several limitations remain. First, the current validation is restricted to $2\times$ magnification and a limited set of lithologies, and the generalization of HLF-SASR to higher magnification factors and more heterogeneous or anisotropic rock types requires further investigation. Second, the present framework operates on 2D slices, whereas many practical applications in digital rock physics demand consistent 3D volume reconstruction and subsequent flow simulation.

Third, although the proposed frequency-aware modeling is implemented via a lightweight feature-space approximation, the incorporation of explicit frequency-domain representations (e.g., wavelet- or Fourier-based decompositions) may provide more principled control over frequency bands and warrants further exploration.

Future work will therefore focus on four directions: extending the core mechanisms of HLF-SASR to a fully 3D SR framework for volumetric micro-CT data to achieve high-fidelity reconstruction of three-dimensional pore networks; incorporating physical constraints such as permeability or flow-related quantities to develop physics-informed SR models that enforce consistency with underlying seepage and transport laws; exploring lightweight, hardware-aware variants suitable for deployment on on-site or edge devices in field-scale digital rock workflows; and investigating the integration of explicit frequency-domain representations, such as wavelet- or Fourier-based decompositions, with deep super-resolution frameworks to further enhance frequency-aware modeling.

In summary, HLF-SASR provides an effective and interpretable technical path for solving the SR reconstruction problem of digital core images. Its core ideas also have reference significance for other scientific imaging tasks with frequency-domain heterogeneity and multi-scale characteristics.

Data Availability

All data supporting the findings of this study are available within the paper and its Supplementary Information. The digital rock image datasets used in this study are publicly accessible. All experiments were conducted on the DRSRD1, introduced by Wang et al. (2019b). DRSRD1 is distributed through the Digital Porous Media portal as part of the Digital Rock Physics project (dataset ID: DRP-211; DOI: 10.17612/P7D38H). The dataset can be accessed via the DPM published-datasets page (`https://digitalporousmedia.org/published-datasets/drp.project.published.DRP-211`).

Competing Interest Declaration

The authors declare no competing interests.

Funding

This work was supported by CNPC Innovation Fund (No.2024DQ02-0501), Royal Society (IEC_NSFC_233444), Youth Science and Technology Talent Promotion Project of Jiangsu Province (JSTJ-2025-137)

Author Contributions

G.T. and Z.C. wrote the main manuscript. L.M prepared Figure. 1 and Figure. 2. N.Z. perform the experiments. X.Y and W.F valided the experiments with analysis. All authors reviewed the manuscript.

References

Chen H, He X, Teng Q, Sheriff R E, Feng J, Xiong S (2020) Super-resolution of real-world rock micro-computed tomography images using cycle-consistent generative adversarial networks. Physical Review E 101(2):023305, ISSN 2470-0045,2470-0053

Chen L C, Papandreou G, Schroff F, Adam H (2017) Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:170605587

Chen L C, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In Computer Vision – ECCV 2018, Springer International Publishing, ISBN 9783030012335,9783030012342, 833–851

Chen P, Shen B, Yang Y, Wu W, Chen S, Zhou G, Malik T, Kalathingal S, Hu J, Tay F R, Ma J (2026) Deep learning super-resolution for dental cbct using micro-ct reference and edge loss function. Journal of Dentistry 164:106209, ISSN 0300-5712

Cnudde V, Boone M (2013) High-resolution x-ray computed tomography in geosciences: A review of the current technology and applications. Earth-Science Reviews 123:1–17, ISSN 0012-8252

Dong C, Loy C C, He K, Tang X (2016) Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 38(2):295–307, ISSN 0162-8828,2160-9292

Fuoli D, Van Gool L, Timofte R (2021) Fourier space losses for efficient perceptual image super-resolution. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2340–2349

Guo S, Yong H, Zhang X, Ma J, Zhang L (2023) Spatial-frequency attention for image denoising. arXiv preprint arXiv:230213598

Ha N Y Y, Ong L Y, Leow M C (2024) Slowfast-tcn: A deep learning approach for visual speech recognition. Emerging Science Journal

Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 7132–7141

Huang H, He R, Sun Z, Tan T (2017) Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 1698–1706

Jang E, Gu S, Poole B (2016) Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:161101144

Jordan M I, Jacobs R A (1994) Hierarchical mixtures of experts and the em algorithm. Neural computation 6(2):181–214

Ledig C, Theis L, Huszar F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, Shi W (2017) Photo-realistic single image super-resolution using a generative adversarial network. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 105–114

Li G, Cui Z, Li M, Han Y, Li T (2024) Multi-attention fusion transformer for single-image super-resolution. Scientific Reports 14(1):10222, ISSN 2045-2322

Li J, Fang F, Mei K, Zhang G (2018) Multi-scale residual network for image super-resolution. In Computer Vision – ECCV 2018, Springer International Publishing, ISBN 9783030012366,9783030012373, 527–542

Lim B, Son S, Kim H, Nah S, Lee K M (2017) Enhanced deep residual networks for single image super-resolution. In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 1132–1140

Liu W, Shang L, Gao G, Li D, Wang M, Geng W, Jia Q, Zhang G, Zhang H, Hao P, Peng L (2025) Microstructure characterization of tight reservoirs using micro-ct and fib-sem imaging technology. Frontiers in Earth Science 13:1551826, ISSN 2296-6463

Niu Y, Jackson S J, Alqahtani N, Mostaghimi P, Armstrong R T (2021) A comparative study of paired versus unpaired deep learning methods for physically enhancing digital rock image resolution. arXiv preprint arXiv:211208644

Otoom M M, Etoom A M (2025) Leveraging image analysis and deep convolutional neural networks for cutting-edge malware detection and mitigation. HighTech and Innovation Journal 6(2):662–686

Ren Y, Zeng C, Li X, Liu X, Hu Y, Su Q, Wang X, Lin Z, Zhou Y, Hu H, et al. (2025) Intelligent evaluation of sandstone rock structure based on a visual large model. Petroleum Exploration and Development 52(2):548–558

Shan L, Liu C, Liu Y, Tu Y, Chilukoti S V, Hei X (2024) Single image multi-scale enhancement for rock micro-ct super-resolution using residual u-net. Applied Computing and Geosciences 22:100165, ISSN 2590-1974

Wang H, Dalton L, Fan M, Guo R, McClure J, Crandall D, Chen C (2022) Deep-learning-based workflow for boundary and small target segmentation in digital rock images using unet++ and ik-ebm. Journal of Petroleum Science and Engineering 215:110596, ISSN 0920-4105

Wang X, Yu K, Wu S, Gu J, Liu Y, Dong C, Qiao Y, Loy C C (2019a) Esrgan: Enhanced super-resolution generative adversarial networks. In Computer Vision – ECCV 2018 Workshops, Springer International Publishing, ISBN 9783030110208,9783030110215, 63–79

Wang Y, Armstrong R, Mostaghimi P (2019b) Diverse super resolution dataset of digital rocks (deeprock-sr): Sandstone, carbonate, and coal. Digital Rocks Portal

Wei P, Xie Z, Lu H, Zhan Z, Ye Q, Zuo W, Lin L (2020) Component divide-and-conquer for real-world image super-resolution. In Computer Vision – ECCV 2020, Springer International Publishing, ISBN 9783030585976,9783030585983, 101–117

Xiao Y, Dong Y (2025) Optimizing aigc technology for iot devices with deep learning. HighTech and Innovation Journal 6(3):976–990

Xing Z, Yao J, Liu L, Sun H (2023) Digital rock resolution enhancement and detail recovery with multi attention neural network

Zhang J, Wang Q, Zhang X, Yin Y (2025a) Lofi: Harnessing attention dynamics for facial expression recognition with noisy labels. In ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, IEEE, 1–5

Zhang Y, Bi J, Xu L, Xiang H, Kong H, Han C (2025b) Lightweight multi-scale attention feature distillation network for super-resolution reconstruction of digital rock images. Geoenergy Science and Engineering 246:213628, ISSN 2949-8910

Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y (2018) Image super-resolution using very deep residual channel attention networks. In Computer Vision – ECCV 2018, Springer International Publishing, ISBN 9783030012335,9783030012342, 294–310

Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 6230–6239

Zhou S, Chen D, Pan J, Shi J, Yang J (2024) Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2952–2963

Zhu C, Wang J, Zhang L, Liang J, Su Q, Li B (2025) Sam-fusenet: Segment anything guided multimodal fusion for rgb–thermal aerial robotic perception. IEEE Transactions on Geoscience and Remote Sensing 64:1–12

Zhu Q, Li P, Li Q (2023) Attention retractable frequency fusion transformer for image super resolution. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 1756–1763