npj | heritage science

Article

https://doi.org/10.1038/s40494-025-01718-7

Ancient mural super-resolution reconstruction based on conditional diffusion model for enhanced visual information

Yunsheng Chen[1,2], Aiwu Zhang[1,2] ✉, Feng Gao[3] ✉ & Juwen Guo[3]

Ancient murals are cultural treasures with high research value, yet few are well preserved. The intricate textures within murals pose substantial challenges for super-resolution reconstruction. To address issues such as detail loss, color shifts, and inadequate noise control in mural super-resolution reconstruction, this study introduces a mural conditional diffusion model (MCDM) for enhanced image reconstruction. The model integrates three core modules: a residual feature distillation network for feature encoding and detail extraction, a residual self-attention module to enhance global consistency, and a Kolmogorov–Arnold-based implicit representation module for high-frequency detail reconstruction. Furthermore, this study establishes three training schemes across two datasets. Experiments show that MCDM performs best on the mixed dataset, with a PSNR of 23.3 dB and an SSIM of 0.8399. Tests across scenarios show that MCDM ensures smooth images while preserving details. Transfer learning further enhances performance by reducing noise and restoring fine features, providing guidance for future work.

Ancient Chinese murals, with their long history and rich artistic heritage, reflect the religious beliefs, social conditions, and advanced painting techniques of their time. However, over time, environmental and human factors have caused various forms of deterioration in murals, including surface peeling, pigment loss, dust accumulation, and scratches, which result in blurred patterns and loss of fine textures, posing a serious threat to their artistic and historical value[1–3]. The digital preservation and transmission of mural images face multiple challenges, including lighting conditions, camera angles, capture device resolution, and image compression algorithms. These factors can result in blurred details, increased noise, and color distortion, further compromising image quality and hampering effective display and cultural dissemination.

To enhance the visual quality and detail presentation of mural images, image reconstruction techniques are particularly important. In recent years, single image super-resolution (SISR) has found extensive applications in diverse computer vision tasks. SISR focuses on recovering high-resolution (HR) images from low-resolution (LR) inputs. However, SISR is inherently an ill-posed problem, where a single LR image can correspond to multiple plausible HR solutions. SISR approaches are broadly classified into three categories: interpolation-based methods, reconstruction-based methods[4,5], and learning-based methods[6–8]. The first two types of methods, constrained by their simplicity, are inadequate for addressing the diversity and complexity of mural images and lack the capability to learn contextual relationships and intricate patterns from large-scale data, thus limiting their practical applications. Recently, research has increasingly focused on learning-based methods[9], which capitalize on data-driven techniques to effectively handle the complexity of mural images and fulfill the demands for detail restoration.

Learning-based approaches primarily comprise convolutional neural networks (CNN), generative adversarial networks (GAN), and flow-based models. Most CNN-based single image super-resolution models commonly employ mean squared error (MSE) or mean absolute error (MAE) as loss functions. While effective in optimizing peak signal-to-noise ratio (PSNR), these pixel-level losses often produce overly smoothed reconstructions, limiting the recovery of high-frequency details and intricate textures[10,11]. To overcome these drawbacks, research efforts have progressively focused on incorporating advanced network architectures and loss functions. In 2015, Dong et al.[12] introduced the first three-layer CNN model, SRCNN, for

[1]Key Laboratory of 3D Information Acquisition and Application, Ministry of Education, Capital Normal University, 100048 Beijing, China. [2]College of Resource Environment and Tourism, Capital Normal University, 100048 Beijing, China. [3]China Academy of Cultural Heritage, 100029 Beijing, China.
✉e-mail: zhangaiwu@cnu.edu.cn; gaofeng@cach.org.cn

super-resolution reconstruction, which marked the beginning of a research surge in learning-based super-resolution methods. Subsequently, researchers leveraged residual networks, channel attention modules, and other innovations to develop deeper CNN models, enabling the extraction of richer image features while mitigating training challenges. To enhance efficiency, ESPCN[13] introduced small filters, EDSR[14] incorporated residual blocks with skip connections, and CARN[15] implemented a recursive network architecture, all of which demonstrated excellent reconstruction performance. Building on the successful application of CNNs in traditional image tasks, Xu et al.[16] introduced a multi-scale residual attention network for mural image super-resolution, effectively reconstructing texture-rich mural images.

With their exceptional ability to generate high-quality reconstructed images, generative adversarial networks (GANs) have gradually established themselves as the dominant framework and backbone for image super-resolution reconstruction. GAN-based models integrate content losses (e.g., L1 or L2) with adversarial losses to generate HR images with enhanced perceptual quality, effectively addressing the issue of over-smoothing[17,18]. Nevertheless, these models are susceptible to mode collapse and frequently encounter challenges in achieving stable convergence during training[19]. In comparison, flow-based methods address the training instability inherent in GAN models through explicit distribution modeling, invertibility, and likelihood maximization techniques. However, these methods are often constrained by high memory requirements and significant computational costs, stemming from the necessity of rigorous model designs, such as multi-layer bidirectional transformations or block-structured architectures[20]. In the domain of mural super-resolution, Cao et al.[21] introduced a self-attention GAN-based reconstruction method to mitigate texture blurring and detail loss. This approach utilizes self-attention modules for feature extraction and sub-pixel convolution layers to generate HR images. Experimental results demonstrated significant improvements in PSNR and (structural similarity index) (SSIM), successfully restoring mural textures and meeting public visual quality expectations. Xiao et al.[9] introduced AM-ESRGAN, combining attention mechanisms with multi-level residual networks to achieve enhanced visual fidelity and improved detail restoration. To overcome challenges such as incomplete detail restoration and image distortion, Ren et al.[22] developed a GAN model that integrates parallel dual-convolution feature extraction and a ternary heterogeneous discriminator, delivering high-quality mural image restoration.

In recent years, diffusion models have found extensive applications in various fields, including speech synthesis[23] and image synthesis[24]. Diffusion models, as probabilistic generative models leveraging Markov chains, employ forward diffusion to incrementally add noise and map data to a Gaussian distribution, while reverse diffusion progressively removes noise to recover the original data distribution. By optimizing the variational lower bound (or its simplified version), diffusion models enable a stable training process, explicitly capturing data distributions and producing high-quality, diverse outputs. In comparison to GANs, diffusion models effectively mitigate the mode collapse issue associated with adversarial training, offering superior stability and enhanced detail recovery. Building on this foundation, researchers have introduced several variants, including diffusion probabilistic models[25], conditional score models[26], and denoising diffusion probabilistic models[27]. Among these, conditional diffusion models integrate conditional information into the generation process, allowing the model to produce samples tailored to specific conditions. SRDiff[28] pioneered the application of diffusion models in single-image super-resolution tasks, achieving diverse and realistic predictions on facial and natural image datasets while effectively addressing challenges like over-smoothing and mode collapse. SR3[29] advanced this approach by leveraging bicubically interpolated LR images as conditional inputs and utilizing stochastic denoising within diffusion models, producing more realistic image results on facial and natural image datasets, outperforming GAN-based methods. Li et al.[30] introduced a dual-conditional diffusion model-based fusion network for producing HR hyperspectral images, effectively addressing blur effects commonly observed in deep learning methods. Drawing inspiration from these advancements, we propose an enhanced diffusion model designed to more effectively recover the intricate details and structures present in murals. Li et al.[28] introduced SRDiff, a diffusion model-based super-resolution approach that pioneered the application of diffusion models to super-resolution tasks, achieving remarkable detail restoration on facial and natural image datasets. Dong et al.[31] developed the ISPDiff model, which combines diffusion models to tackle the challenges of hyperspectral images, including their physical properties and inference efficiency, resulting in substantial improvements in spectral fidelity and detail quality. Moreover, diffusion models have been utilized in medical image enhancement, particularly in the super-resolution reconstruction of ultrasound and magnetic resonance imaging (MRI), producing high-quality images with well-defined structures and intricate textures, while demonstrating superior performance on metrics like PSNR and SSIM[32,33].

In conclusion, to address the challenges of over-smoothing and mode collapse in traditional SISR models, while capitalizing on the superior detail restoration capabilities of diffusion models, this study pioneers the application of diffusion models in mural image super-resolution reconstruction. The aim is to effectively preserve and restore the intricate details and colors of mural images, offering technological support for their digital preservation and cultural heritage.

In summary, the main contributions of this paper are as follows:

(1) This paper proposes a mural images super-resolution network based on the diffusion model. By leveraging features extracted from LR images as conditional guidance, the network accurately predicts the noise distribution, thereby reconstructing high-fidelity mural images with rich details from noisy inputs.
(2) This study introduces a novel implicit neural representation module grounded in the Kolmogorov–Arnold network (KAN), designed to replace the conventional multilayer perceptron module. Leveraging a limited set of basis functions, this module efficiently approximates complex high-dimensional data mappings, aiding in the capture of nonlinear image features and preventing excessive smoothing, thus improving detail reconstruction.
(3) This study introduces a dynamic feature weighting module based on residual connections and self-attention mechanisms (RSAM), which adaptively adjusts feature weights during downsampling, leveraging self-attention to capture long-range dependencies and maintain global structural consistency. RSAM accentuates critical features, suppresses irrelevant information, effectively alleviates the vanishing gradient problem, and strengthens information propagation within deep networks.
(4) This study incorporates a residual feature distillation network (RFDN) to efficiently capture and refine both local details and global structural features in mural images. The RFDN incrementally distills crucial features, extracting essential texture and color information from LR images to enhance the restoration of intricate details, thereby improving the quality of HR mural image reconstruction.

Methods

Overall structure

The overall network architecture proposed in this study is shown in Fig. 1. Since the diffusion model employs an iterative solution process, the output of MCDM is used as input for the next step at each iteration, continuing until a predetermined number of steps is completed. MCDM takes as input an LR digital mural image and the random noise from the previous iteration. Specifically, the LR image is first fed into the condition-embedding layer, where it undergoes feature encoding and refinement via the embedded RFDN to extract detailed conditional features. Subsequently, these features are passed to the DN-Module on the left, which is built upon the StyleGAN module from StyleGAN2[34] to further extract and refine image features.

During feature extraction, the encoded features are concatenated with random noise and fed into the DNS-Module, composed of the RSAM, to capture long-range dependencies in the image. Both the DN-Module and DNS-Module propagate gradients during the encoding phase, passing

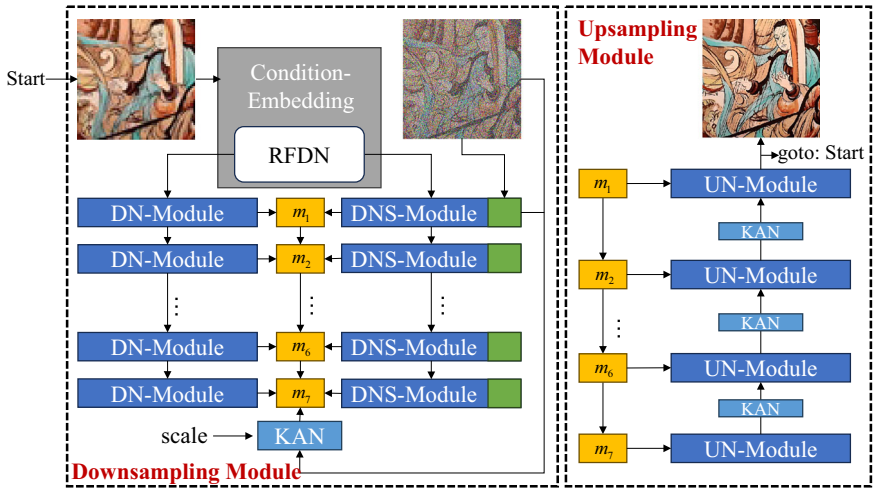

Fig. 1 | The overall structure of the MCDM. This figure was created using Microsoft PowerPoint.

intermediate results to the subsequent m_n layer. Each m_n layer is composed of RSAM, with the encoded results utilized during the up-sampling phase for recovering image details and gradients propagated downward during encoding.

In a distinctive design of the encoding process, the noise and the encoded output from the first DNS-Module are simultaneously fed into the KAN layer, accompanied by a scale parameter. The KAN layer processes the input features further and passes the resulting output to the final layer of the m_n layer, enhancing the model's capability to represent fine details in image reconstruction. Subsequently, these processed features are passed to the up-sampling module for further enhancement.

During the up-sampling stage, features from each m_n layer are sequentially fed into the UN-Module. Comprising the fundamental components of StyleGAN2, the UN-Module alternates with the KAN layer to perform up-sampling steps, gradually reconstructing an HR image.

Conditional diffusion model

The conditional diffusion model comprises two primary stages: a forward noise addition (diffusion) process and a reverse denoising process. By incrementally adding Gaussian noise to the original image and subsequently removing it during the reverse phase.

During the forward diffusion phase, the model defines a forward Markov diffusion process q, incrementally adding Gaussian noise to the HR original image x_0, gradually transforming the image into a pure noise distribution. This process adds noise incrementally over T time steps, forming a Markov chain, with its conditional probability can be expressed as

$$q(x_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(x_t|\boldsymbol{x}_{t-1}) \tag{1}$$

$$q(x_t|x_{t-1}) = \mathrm{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\boldsymbol{I}) \tag{2}$$

The hyperparameter α_t controls the noise intensity added at each iteration, where $0 < \alpha_t < 1$ and $\boldsymbol{I}$ is the identity matrix. Since standard Gaussian noise is added at each step, this process can be considered a Markov process. Leveraging this property, the noisy image x_t at any arbitrary time step can be directly computed from the original HR image x_0 using the following equation:

$$x_t = \sqrt{\gamma}\boldsymbol{x}_0 + \sqrt{1-\gamma}\epsilon \tag{3}$$

where $\gamma = \prod_{i=1}^{t} \alpha_i$, ϵ is a noise term drawn from a standard normal distribution. This formulation enables direct computation of the noisy image at any given time step, bypassing the need for iterative calculations at each step.

During the reverse denoising phase, the model gradually removes noise via a reverse Markov chain, ultimately yielding a high-quality image. Starting from a high-noise state x_t,the model leverages the conditional probability $p(x_{t-1}|x_t)$ to iteratively recover a high-quality image that approximates the original. At each time step t, the model predicts the image of the previous state x_{t-1}, continuing until it generates a low-noise or noise-free image x_0. The reverse process is formulated as the following difference equation:

$$x_{t-1} = \boldsymbol{\mu}_\theta(x_t, t) + \sigma_t \cdot \epsilon \tag{4}$$

where $\boldsymbol{\mu}_\theta(x_t, t)$ denotes the mean function estimated by the model, conditioned on the current state x_t and time step t,while σ_t represents the time-dependent standard deviation.

Implicit neural representation based on KAN

Traditional implicit neural representations using multilayer perceptron (MLP) are effective at capturing complex features, but high-dimensional feature representation often demands extensive computational resources and deeper network architectures, leading to increased computational complexity. According to the Kolmogorov–Arnold theorem, any high-dimensional nonlinear function can be approximated with appropriately selected low-dimensional basis functions (e.g., polynomials or splines), thus mitigating the computational complexity inherent in layer-by-layer approximation used in traditional neural networks like MLP[35,36].

Motivated by this insight, we propose an innovative implicit neural representation module based on the Kolmogorov–Arnold network, intended to replace conventional MLP modules. This module leverages a minimal number of basis functions to efficiently represent complex features in mural images, reducing computational complexity and enhancing reconstruction accuracy. In this work, spline functions are employed as basis functions to construct an implicit neural representation module grounded in the Kolmogorov–Arnold network. This module effectively captures complex high-dimensional data mappings and integrates with the diffusion model, forming a unified end-to-end framework. In the image decoding phase, the diffusion model incrementally removes noise, leveraging the feature representation capabilities of the implicit neural representation module to reconstruct a HR mural image. The specific process resembles the LIIF model[37], where continuous coordinates of multi-resolution features $c = \{c(1), c(2), \ldots, c(N)\}$ serve as references, with their values retrieved from the noise network via a scaling factor. We input the current features surrounding the coordinates to calculate the target features. Given the feature $h^{(i+1)}$ and its corresponding coordinates $c^{(i+1)}$, The implicit representation process draws on the principles of the IDM framework[38]. the implicit

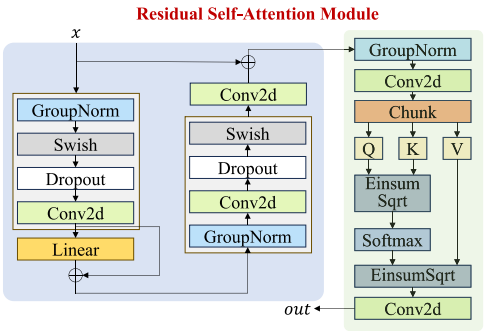

Fig. 2 | The structure of the RSAM. This figure was created using Microsoft PowerPoint.

representation process can be expressed as follows:

$$\mathbf{n}^{(i)} = K_i\left(\hat{\mathbf{h}}^{(i+1)}, c^{(i)} - \hat{c}^{(i+1)}\right) \tag{5}$$

where c denotes the continuous coordinates of multi-scale images. In the $(i+1)$-*th* depth, $\hat{\mathbf{h}}^{(i+1)}$ and $\hat{c}^{(i+1)}$ are interpolated by computing the nearest Euclidean distance from $\mathbf{h}^{(i+1)}$ and $c^{(i+1)}$, respectively. K_i refers to KAN[39]. The structure of KAN can be represented by the following equation:

$$f(x) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \varphi_{q,p}(x_p) \right) \tag{6}$$

The function $f(x)$ is the output, with its input being a multidimensional vector $x = (x_1, x_2, \ldots, x_n)$,where $x_1, x_2, \ldots, x_n$ are distinct independent variables.Φ_q represents a family of functions, each of which operates on the result of an inner summation. The function $\Phi_q : \mathbb{R} \to \mathbb{R}$ can be understood as a transformation applied to the result of each inner summation. The index q ranges from 1 to $2n+1$, indicating that there are $2n+1$ such transformation functions. $\varphi_{q,p}$ is another family of functions, each of which operates on a single independent variable x_p. Each function $\varphi_{q,p} : [0,1] \to \mathbb{R}$ is a transformation function applied to the individual variable x_p, yielding a real-valued result. The index p ranges from 1 to n, indicating that there are n input variables, with each input variable x_p having a corresponding transformation function.

Residual self-attention dynamic weighting module

The RSAM constitutes an essential component of the MCDM model, incorporating two core elements: residual connections and self-attention mechanism, as shown in Fig. 2.

The residual connections module on the left side processes input features via identity mapping, preserving the original data information while directly passing error signals, thus effectively preventing gradient vanishing during deep network training. Initially, input features are normalized using the GroupNorm module[40]. The normalized output is subsequently fed into the Swish activation function, which helps prevent the gradient from approaching zero during training, thereby mitigating saturation effects. A Dropout layer is then applied to introduce model uncertainty and reduce complexity, thereby enhancing generalization capacity. Finally, the output passes through the last convolutional layer, is fed into a Linear layer, and undergoes residual computation. The residual output is fed back into the network, where the above processes are iterated.

The self-attention mechanism module on the right dynamically adjusts feature weights based on relationships among input features, emphasizing the most representative and significant features while attenuating irrelevant or noisy ones. This module, integrating residual connections with self-attention, significantly improves the model's capacity to understand input data, enhancing generalization and accuracy on complex datasets. Upon

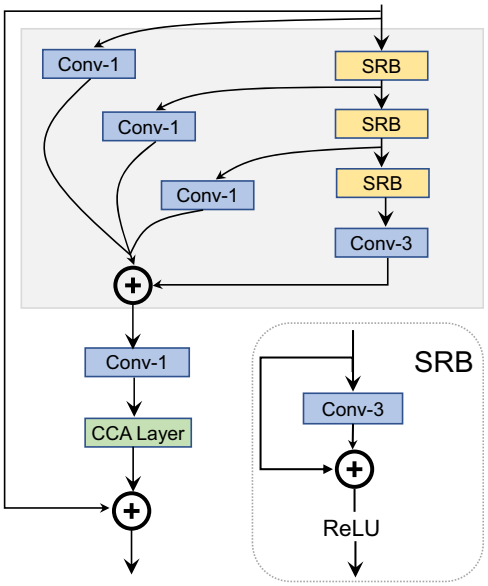

Fig. 3 | The main structure of the RFDN. This figure was created using Microsoft PowerPoint.

entering the self-attention module on the right, the data undergoes GroupNorm normalization. Subsequently, after a convolution operation, the features are divided into three tensor chunks using Torch's built-in Chunk function, representing the query (Q), key (K), and value (V). The query (Q) and key (K) are utilized to compute the attention scores, The equation can be expressed as follows:

$$A[b,n,h,w,y,x] = \frac{1}{\sqrt{c}} \sum_{j=1}^{c} Q[b,n,j,h,w] \cdot K[b,n,j,y,x] \tag{7}$$

where b denotes the batch index, n represents the attention head index, h and w serve as spatial indices for the Query Matrix, while y and x are spatial indices for the Key Matrix. The channel index c is applied to scale the results, mitigating potential numerical instability or overflow in computations.

The module first computes the attention score A by performing a dot product between the query vector Q at each h, w position and the key vector K at each y, x position, normalized by the square root of the channel dimension. the attention score is then activated using Softmax function, yielding a probability distribution that serves as weights. Subsequently, the weights are applied to the value vector V through a weighted summation to obtain the self-attention score. Finally, the self-attention scores pass through a convolutional layer to further extract deep features, yielding the final attention output.

Residual feature distillation network

To improve feature representation efficiency and extract more effective features for producing high-quality super-resolution images, this study employs the residual feature distillation network (RFDN) for the mural image super-resolution task[41]. The main structure of the network is shown in Fig. 3. Unlike conventional methods that use a fixed 3×3 convolutional kernel for channel compression, the distillation module on the left side of the model replaces the 3×3 kernel with a 1×1 convolutional kernel. This design reduces the parameter count while enabling more efficient extraction of fine details, which is especially important for the rich and intricate textures of mural images.

To better account for the spatial contextual information in mural images, the right main network replaces the original 3×3 convolutional kernel with a refinement module composed of shallow residual blocks (SRB), enhancing the network's ability to extract fine details. This architectural choice enables RFDN to efficiently leverage spatial information, enhance attention computation efficiency, and reduce network complexity,

thus achieving better reconstruction performance in mural super-resolution tasks with a lightweight structure.

Loss function

Given that diffusion models often generate complex and diverse data, and L1 loss does not square errors, making it less sensitive to outliers, this study utilizes L1 loss to mitigate the impact of extreme error values on the overall loss, thereby stabilizing the training process. Additionally, L1 loss and L1 regularization have inherent sparsity properties, which are particularly beneficial for tasks like image denoising and restoration. The use of L1 loss encourages the model to produce a sparse error distribution, effectively eliminating noise and restoring the original structure of the image. The L1 loss function applied in this study is defined as follows:

$$\ell(x, y) = \mathrm{sum}(L) = \sum_{n=1}^{N} |x_n - y_n| \tag{8}$$

where N denotes the batch size, x represents the pixel values of the target image, and y denotes the pixel values of the generated image.

Datasets

In this study, we constructed a mural dataset consisting of 2281 real mural images from the Han, Tang, and other dynasties, as well as reproductions from ancient mural collections. The dataset spans different dynasties, styles, and regions, ensuring both diversity and representativeness of the data. A total of 2031 images were used for training, 100 for validation, and 150 for testing. Additionally, the widely used DIV2K[34] dataset was employed as a benchmark for image super-resolution research. DIV2K comprises a training set of 900 images and a validation set of 100 images, encompassing diverse scenes such as people, natural landscapes, and urban environments. This dataset is suitable for testing, benchmarking, and pre-training super-resolution algorithms.

Experimental design

The hardware configuration for the experiments includes 64 GB of RAM, an Intel Core i9-12900KF CPU, and an NVIDIA GeForce 4090 GPU. The software environment was built on an Ubuntu operating system running the Linux kernel, with Python 3.9 as the programming language. Key libraries used include PyTorch, Numpy, and OpenCV-Python.

In the data preprocessing stage, random noise and blur inherent to mural images were simulated by downsampling HR images. Specifically, the Lanczos algorithm[42] was employed to downscale the images by a factor of 8, creating LR samples. Both HR and LR images were then cropped to the largest inscribed square, with the initial cropping position randomly selected to enhance sample diversity. The model was trained using the Adam optimizer, with a learning rate of 1×10^{-4} and a total of 1×10^{6} iterations.

In this study, two datasets were used, and three training schemes were designed. In the first scheme, the mural dataset was used as the training set. The second scheme combined the mural dataset with the DIV2K dataset to create a mixed dataset. In the third scheme, the model was pre-trained on the DIV2K dataset, and then the parameters were transferred to the mural dataset for fine-tuning.

To assess the effectiveness of the proposed mural image restoration model, evaluations were conducted on a self-constructed test dataset. Both quantitative and qualitative analyses were conducted to compare the proposed model with existing super-resolution algorithms (EDSR[19], SRGAN[14], CARN[15], LKDN[43], and SRCNN[12]), evaluating their performance in the mural image restoration task.

Evaluation metrics

In the super-resolution reconstruction of digital mural images, the performance is primarily assessed using two essential metrics: PSNR and SSIM. These metrics are essential for analyzing and enhancing the restoration quality of mural images, particularly focusing on detail preservation and visual coherence. These metrics provide an effective means of comparing different super-resolution algorithms, specifically in terms of improving mural image clarity and detail fidelity.

PSNR is derived from the mean squared error (MSE) metric. For an HR image X and a SR image Y of dimensions $m \times n$, MSE is defined as

$$\mathrm{MSE} = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [X(i,j) - Y(i,j)]^2 \tag{9}$$

Derived from the MSE, PSNR can be calculated as follows, where MAX_{Image} represents the maximum gray value in the images.

$$\mathrm{PSNR} = 10 * \log_{10}\left(\frac{\mathrm{MAX}_{\mathrm{Image}}^2}{\mathrm{MSE}}\right) \tag{10}$$

SSIM is commonly used to quantify the similarity between images before and after distortion, providing a measure of the realism of model-generated images. SSIM is defined as

$$SSIM(X, Y) = \frac{(2\mu_X\mu_Y + c_1)(2\sigma_{XY} + c_2)}{(\mu_X^2 + \mu_Y^2 + c_1)(\sigma_X^2 + \sigma_Y^2 + c_2)} \tag{11}$$

where μ_X and μ_Y are the mean values of X and Y, respectively. σ_X^2 and σ_Y^2 are the variances of X and Y, respectively. σ_{XY} is the covariance between X and Y. c_1 and c_2 are small constants, proportional to the dynamic range of the image, introduced to prevent division by zero.

Results

Comparative analysis of single dataset training using the mural dataset

In this study, three representative digital mural reconstruction results were selected, corresponding to single-character, multi-character, and complex scenes (as shown in Fig. 4). The detailed analysis is as follows:

In the single-character scene (Scene 1), CARN, LKDN, SRCNN, and our model effectively restore the mural's overall color palette. However, EDSR exhibits severe color shifts, SRGAN struggles with skin and background restoration, rendering a cyan tone, and ESPCN demonstrates missing background details with a blue color bias. Although SRCNN restored the color, it lacked sufficient detail in the outlines. CARN and LKDN effectively preserved the character's contours but missed finer details in the hands and exhibited more noise. In contrast, our model accurately restored the character's lines and colors.

In the multi-character scene (Scene 2), EDSR exhibits a color shift, resulting in an overall brownish tone. CARN, LKDN, and SRCNN restore the sharpness of handheld objects effectively, while ESPCN and SRGAN struggle to recover finer details. Our model achieves accurate restoration for most outlines, excluding the eyes and neck.

In the complex scene (Scene 3), involving characters with intricate movements and additional animals, EDSR demonstrates significant color distortion, while ESPCN suffers from poor edge restoration and noticeable noise. Although CARN, LKDN, SRGAN, SRCNN, and our model successfully restore the black snake in the image, none of the models adequately capture the facial and clothing details of the character.

As depicted in Fig. 5, we analyzed complex details from three scenes. In Scene 1, all models except ours failed to reconstruct the contour lines and showed noticeable noise. Although our model achieves good sharpness along the edges, its performance in restoring the eye details is suboptimal. In Scene 2, the other models' results contain significant noise, while our model better restores certain contours. Nevertheless, it struggles with the accurate reconstruction of small, elongated objects like prayer beads and flutes. In Scene 3, our model successfully restores the colors and contours of both the snake and the character, although it underperforms in capturing finer facial details. CARN, LKDN, and SRCNN also manage to reconstruct the snake's colors and contours effectively, but display high noise in the character's details.

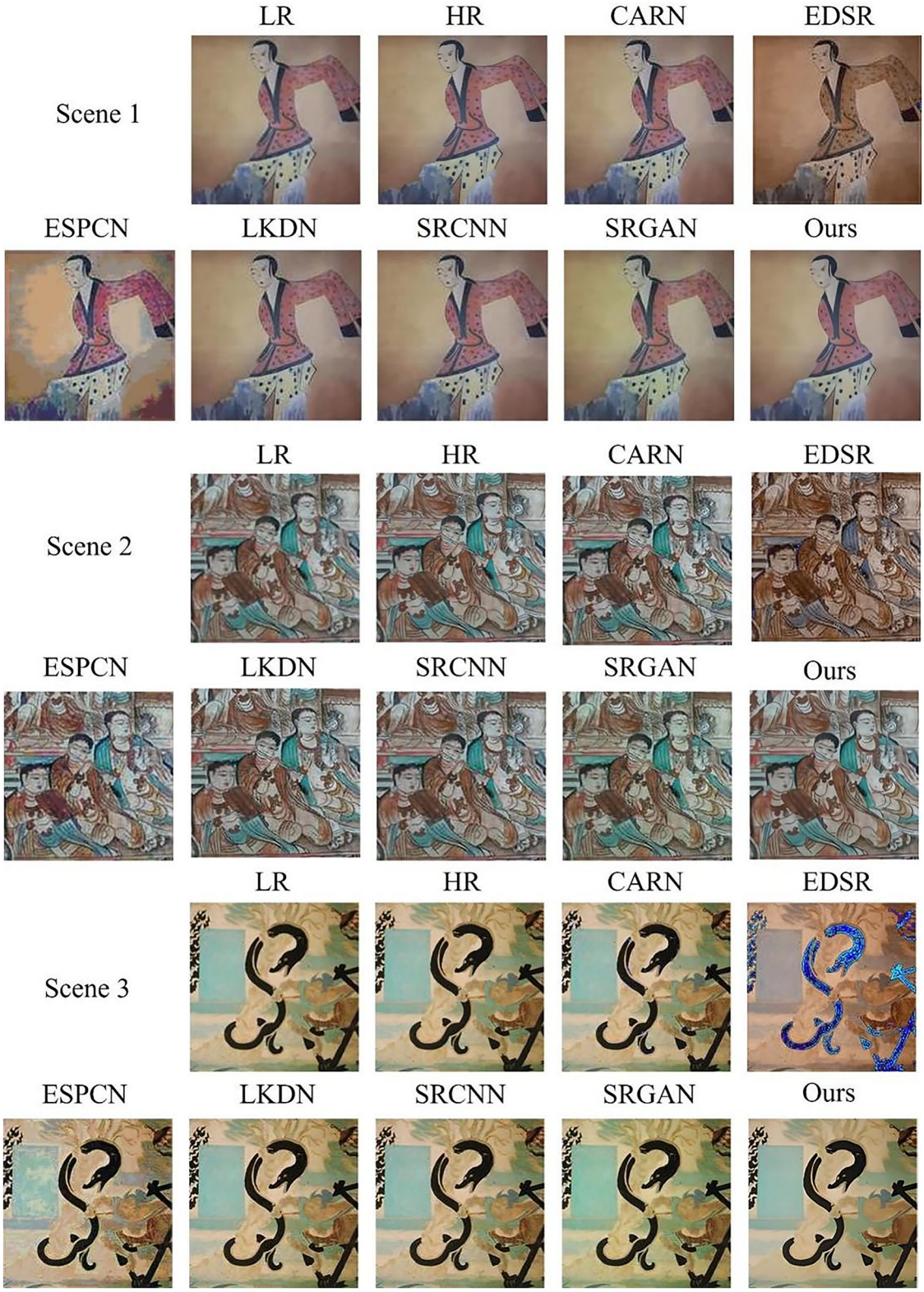

Fig. 4 | Results of various models on the mural dataset. The data in this image comes from the results generated by the model.

Comparative analysis of mixed dataset training with mural and DIV2K datasets

Figure 6 presents the training performance of different models on the Mixed Dataset. EDSR, in particular, fails to accurately restore the overall color across multiple scenes, displaying a noticeable purple color shift. In Scene 1, CARN and LKDN successfully reconstruct the character's body contours with minimal noise, whereas SRGAN and SRCNN exhibit a moderate level of noise, and ESPCN produces the highest noise levels among all models. Our model demonstrates superior noise control, accurately restoring the contours of the character's skin, clothing, and background details with clarity. In Scene 2, CARN, LKDN, and our model achieve effective noise reduction, whereas other models still exhibit noticeable noise levels. In Scene

Fig. 5 | Detailed restoration results of various models on the mural dataset. The data in this image comes from the results generated by the model.

3, SRCNN exhibits blurred edges in the restoration of branches, whereas all models, with the exception of ESPCN and EDSR, achieve near-original quality in reconstructing the ultra-low-resolution complex scene.

A magnified analysis of the details in Fig. 6 is presented in Fig. 7. In Scene 1, CARN, LKDN, and our model accurately restored the facial contours of the character. However, while CARN and LKDN produced somewhat blurred eye details, our model achieved a more refined restoration in this area. SRCNN and SRGAN achieved a certain level of noise reduction, with SRGAN demonstrating smoother results than SRCNN. Although ESPCN successfully restored the overall color tone, it introduced noticeable noise in the facial area. In the multi-character complex scene (Scene 2), CARN and LKDN achieved effective noise reduction and accurately restored neck contours; however, the generated images lacked high-frequency details, resulting in blurred color patches. Our model

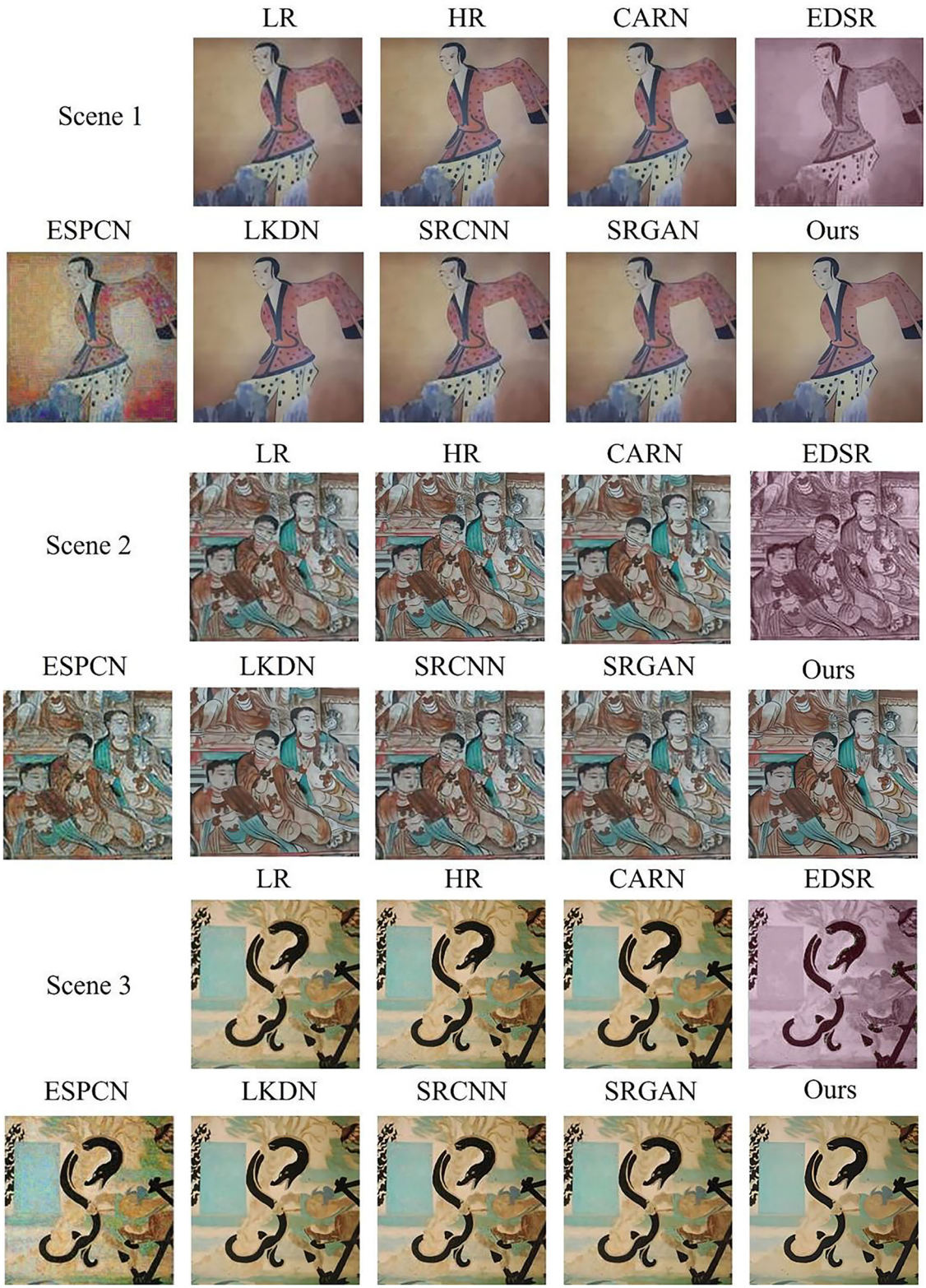

Fig. 6 | Performance of various models trained on the mixed dataset. The data in this image comes from the results generated by the model.

accurately restored the contours of clothing and preserved high-frequency details more effectively. Qualitatively, our model demonstrated superior performance in restoring finer details, such as hand and necklace elements of the characters, compared to other models. In Scene 3, CARN, LKDN, SRGAN, and our model all effectively restored the shape and color of the snake; however, our model additionally captured some facial details of the character that were missed by other models.

Comparative analysis of transfer learning: pre-training on DIV2K and fine-tuning on mural dataset

In this study, all models were first pretrained on the public DIV2K dataset, followed by parameter transfer to the mural dataset for further fine-tuning, as shown in Fig. 8. Compared to the mixed dataset training results shown in Fig. 6, the overall color differences are minimal. Although EDSR exhibits higher sharpness in capturing high-frequency details, it still fails to

Fig. 7 | Detailed results of various models trained on the mixed dataset. The data in this image comes from the results generated by the model.

accurately restore the overall color. CARN, LKDN, SRCNN, and our proposed model demonstrate reduced noise compared to the results trained on the mixed dataset. While ESPCN achieves higher sharpness, it still exhibits a considerable amount of noise.

The detailed reconstruction results based on the pretrained model are illustrated in Fig. 9. In Scene 1, CARN, LKDN, and our model all exhibit good sharpness; however, our model provides superior clarity in restoring clothing pattern details. In Scene 2, the pretrained model results display slightly less detail in the hand region compared to the mixed dataset training results (Fig. 7). Furthermore, compared to the results of direct training on the mural dataset (Fig. 5), all methods, except ESPCN, exhibited notable noise reduction across the three scenes and achieved varying degrees of enhancement in detail and color restoration, validating the effectiveness of transfer learning in mural super-resolution tasks.

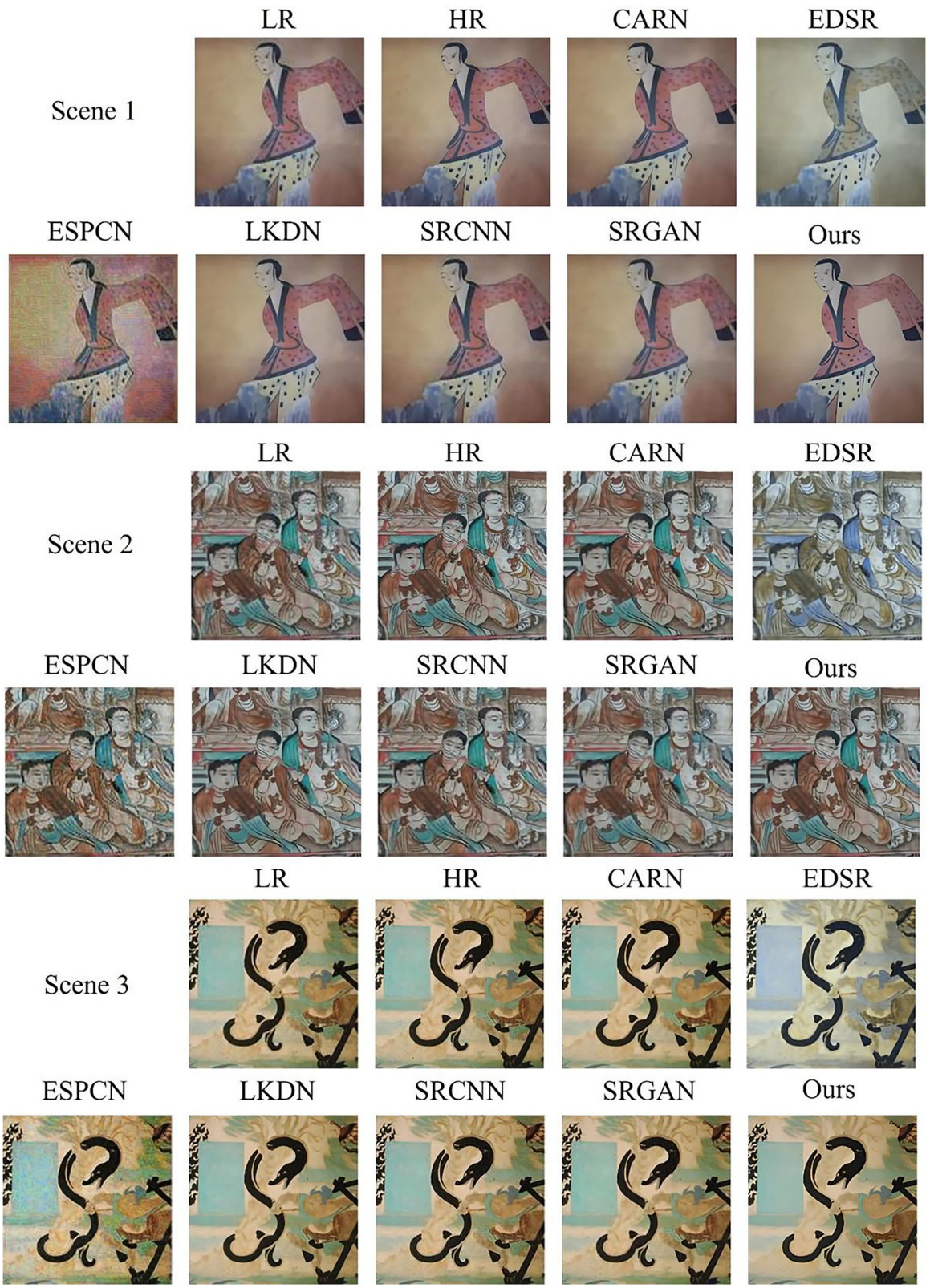

Fig. 8 | Training results of various models based on the pre-trained models. The data in this image comes from the results generated by the model.

Quantitative analysis

Table 1 presents the PSNR and SSIM performance of seven algorithms for image reconstruction tasks across three different training configurations. The MCDM model consistently achieves the highest scores in all training conditions, surpassing the second-best model, LKDN, by 1.1, 0.12, and 0.16 in PSNR, respectively. Additionally, MCDM maintains higher SSIM values, demonstrating its superior ability to model mural image features and its stronger generalization performance. In the mixed dataset training, CARN and LKDN also show promising results; however, under the transfer learning condition, MCDM continues to outperform other models,

Fig. 9 | Detailed display of the training results of various models based on pre-trained models. The data in this image comes from the results generated by the model.

highlighting its strong adaptability. Both SRCNN and LKDN exhibit competitive performance under transfer learning as well. Overall, MCDM demonstrates outstanding performance across various training scenarios, with excellent stability and transferability, making it highly suitable for mural image super-resolution reconstruction in complex data environments.

Ablation experiment

Table 2 presents the ablation study results of individual modules in the mural super-resolution task. Initially, substituting the KAN module with a traditional MLP network led to a 0.67 drop in PSNR and a 0.0679 reduction in SSIM. Keeping the other modules unchanged, removing RFDN caused a

Table 1 | Comparative performance analysis of models on different datasets

Model	Performance on mural dataset		Performance on mixed dataset		Performance after transfer learning on mural dataset	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SRCNN	22.02	0.7185	22.31	0.7325	22.32	0.7341
ESPCN	20.52	0.6890	20.75	0.6886	20.91	0.6890
CARN	20.71	0.7136	22.62	0.7566	21.91	0.7512
SRGAN	21.84	0.7204	22.57	0.7419	22.20	0.7356
LKDN	22.13	0.7268	23.18	0.7639	23.04	0.7597
EDSR	15.98	0.6566	16.14	0.6744	16.88	0.6911
MCDM (ours)	**23.23**	**0.7657**	**23.30**	**0.8399**	**23.20**	**0.8336**

The bold values indicate the best performance for each corresponding metric.

Table 2 | The evaluation results of ablation studies

KAN module	RFDN module	RSAM module	PSNR	SSIM
✗	✓	✓	22.63	0.7720
✓	✗	✓	23.02	0.7902
✓	✓	✗	**22.22**	**0.7499**
✓	✓	✓	**23.3**	**0.8399**

The bold numbers indicate the highest and lowest scores. '✗' denotes the removal of the module, and '✓' denotes its retention.

0.28 decline in PSNR and a 0.0497 decrease in SSIM. Subsequently, removing RSAM resulted in a 1.08 drop in PSNR and a 0.09 decrease in SSIM. Experimental results demonstrate that RSAM has the most substantial influence on model performance, as it effectively captures global information and long-range dependencies, contributing to improved detail restoration and texture generation. In contrast, while replacing the KAN module resulted in a performance decline, its impact was comparatively minor, suggesting that its primary function lies in optimizing local details. Replacing RFDN had the smallest effect on PSNR and SSIM, highlighting its strength in feature extraction; however, its overall contribution was less substantial compared to RSAM and KAN modules.

Discussion

In this study, we propose a mural image super-resolution reconstruction network (MCDM) based on a conditional diffusion model. The MCDM incorporates the KAN-based implicit neural representation, RFDN, and RSAM to accurately restore complex textures and high-frequency details in mural images. By utilizing a progressive detail refinement strategy, MCDM effectively simulates the transformation from LR to HR images, exhibiting notable noise reduction capabilities and superior detail preservation across diverse scenarios.

The ablation study further validates the importance of each module in enhancing reconstruction performance, with the RSAM module showing a particularly prominent contribution in capturing long-range dependencies and improving global structural coherence. Analysis of results across three training schemes shows that MCDM achieves optimal performance on the mixed dataset (PSNR 23.3 dB, SSIM 0.8399), further underscoring its superiority in mural image restoration.

While the MCDM method demonstrates outstanding performance in detail recovery, it still leaves room for improvement in reconstructing certain details within complex scenes. This study confirms the critical potential of transfer learning in mural super-resolution tasks. Future research will delve into dataset development, training strategy refinement, and model parameter optimization to achieve further improvements in mural reconstruction quality.

Data availability

The data presented in this study are available on request from the corresponding author.

Received: 20 November 2024; Accepted: 15 April 2025;
Published online: 22 May 2025

References

1. Pan, X., Ge, Q. & Pan, J. Damage to ancient mural paintings and petroglyphs caused by *Pseudonocardia* sp.—a review. *Acta Microbiol. Sin.* **55**, 813–818 (2015).
2. Veneranda, M. et al. In-situ multianalytical approach to analyze and compare the degradation pathways jeopardizing two murals exposed to different environments (Ariadne House, Pompeii, Italy). *Spectrochim. Acta Part A* **203**, 201–209 (2018).
3. Ogura, D. et al. Influence of environmental factors on deterioration of mural paintings in Mogao Cave 285, Dunhuang. *Case Stud. Build. Rehabil.* **13**, 105–159 (2021).
4. Xu, Y., Li, J., Song, H. & Du, L. Single-image super-resolution using panchromatic gradient prior and variational model. *Math. Probl. Eng.* **2021**, 9944385 (2021).
5. Huang, Y. et al. Single image super-resolution via multiple mixture prior models. *IEEE Trans. Image Process.* **27**, 5904–5917 (2018).
6. Kim, J., Lee, J. K. & Lee, K. M. Accurate image super-resolution using very deep convolutional networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 1646–1654 (IEEE, 2016).
7. Zhang, D., Shao, J., Li, X. & Shen, H. Remote sensing image super-resolution via mixed high-order attention network. *IEEE Trans. Geosci. Remote Sens*. **59**, 5183–5196 (2020).
8. Dai, T. et al. Second-order attention network for single image super-resolution. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Los Angeles, CA, USA* 11065–11074 (IEEE, 2019).
9. Xiao, C. et al. AM-ESRGAN: super-resolution reconstruction of ancient murals based on attention mechanism and multi-level residual network. *Electronics* **13**, 3142 (2024).
10. Zhang, Y. et al. Image super-resolution using very deep residual channel attention networks. In *Proc. European Conference on Computer Vision (ECCV)* (eds Ferrari, V., Hebert, M. Sminchisescu, C. & Weiss, Y.) 286–301 (Springer, 2018).
11. Guo, Y. et al. Closed-loop matters: dual regression networks for single image super-resolution. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 5407–5416 (IEEE, 2020).
12. Dong, C., Loy, C. C., He, K. & Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**, 295–307 (2015).
13. Shi, W. et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 1874–1883 (IEEE, 2016).
14. Lim, B. et al. Enhanced deep residual networks for single image super-resolution. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops* 136–144 (IEEE, 2017).
15. Namhyuk, A., Byungkon, K. & And Kyung-Ah, S. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proc. European Conference on Computer Vision (ECCV)* (eds Ferrari, V., Hebert, M. Sminchisescu, C. & Weiss, Y.) 252–268 (Springer, 2018).
16. Xu, Z., Yan, J. & Zhu, H. Mural image super resolution econstruction based on multi-scale residual attention network. *Laser Optoelectron. Prog.* **57**, 152–159 (2020).
17. Rad, M. S. et al. Srobb: targeted perceptual loss for single image super-resolution. In *Proc. IEEE/CVF International Conference on Computer Vision* 2710–2719 (IEEE, 2019).
18. Zhang, W., Liu, Y., Dong, C. & Qiao, Y. Ranksrgan: generative adversarial networks with ranker for image super-resolution. In *Proc. IEEE/CVF International Conference on Computer Vision* 3096–3105 (IEEE, 2019).

19. Ledig, C. et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 4681–4690 (IEEE, 2017).
20. Lugmayr, A., Danelljan, M., Van Gool, L. & Timofte, R. Srflow: Learning the super-resolution space with normalizing flow. In *Computer Vision —ECCV 2020*. *Lecture Notes in Computer Science* 12360 (eds Vedaldi, A., Bischof, H., Brox, T. & Frahm, J. M.) (Springer, Cham, 2020).
21. Cao, J., Jia, Y., Yan, M. & Tian, X. Superresolution reconstruction method for ancient murals based on the stable enhanced generative adversarial network. *EURASIP J. Image Video Process.* **2021**, 1–23 (2021).
22. Ren, H., Sun, K., Zhao, F. & Zhu, X. Dunhuang murals image restoration method based on generative adversarial network. *Herit. Sci.* **12**, 39 (2024).
23. Kong, Z. et al. Diffwave: a versatile diffusion model for audio synthesis. In *Proc.The 9th International Conference on Learning Representations (ICLR)* (OpenReview, 2021).
24. Yue, Q. et al. Denoising diffusion probabilistic model for face sketch-to-photo synthesis. *IEEE Trans. Circuits Syst. Video Technol.* **34**, 10424–10436 (2024).
25. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. & Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proc. International Conference on Machine Learning (ICML)* (eds Bach, F. & Blei, D.) 2256–2265 (PMLR, 2015).
26. Batzolis, G., Stanczuk, J., Schönlieb, C.-B. & Etmann, C. Conditional image generation with score-based diffusion models. arXiv preprint arXiv:2111.13606. (2021).
27. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *NeurIPS* **33**, 6840–6851 (2020).
28. Li, H. et al. Srdiff: single image super-resolution with diffusion probabilistic models. *Neurocomputing* **479**, 47–59 (2022).
29. Saharia, C. et al. Image super-resolution via iterative refinement. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 4713–4726 (2022).
30. Li, S., Li, S. & Zhang, L. Hyperspectral and panchromatic images fusion based on the dual conditional diffusion models. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–15 (2023).
31. Dong, W. et al. ISPDiff: Interpretable Scale-Propelled Diffusion model for hyperspectral image super-resolution. *IEEE Trans. Geosci. Remote Sens.* **62**, 3407967 (2024).
32. Liu, T. et al. Super-resolution reconstruction of ultrasound image using a modified diffusion model. *Phys. Med. Biol.* **69**, 125026 (2024).
33. Wu, Z. et al. Super-resolution of brain MRI images based on denoising diffusion probabilistic model. *Biomed. Signal Process. Control.* **85**, 104901 (2023).
34. Agustsson, E. & Timofte, R. NTIRE 2017 challenge on single image super-resolution: dataset and study. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops* 126–135 (IEEE, 2017).
35. Vaswani, A. Attention is all you need. *NeurIPS* **30**, 5998–6008 (2017).
36. Cunningham, H. et al. Sparse autoencoders find highly interpretable features in language models. arXiv preprint arXiv:2309.08600 (2023).
37. Chen, Y., Liu, S. & Wang, X. Learning continuous image representation with local implicit image function. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 8628–8638 (IEEE, 2021).
38. Gao, S. et al. Implicit diffusion models for continuous super-resolution. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 10021–10030 (IEEE, 2023).
39. Liu, Z. et al. KAN: Kolmogorov–Arnold networks. arXiv preprint arXiv:2404.19756 (2024).
40. Wu, Y. & He, K. Group normalization. In *Proc. European Conference on Computer Vision (ECCV)* (eds Ferrari, V. & Hebert, M., Sminchisescu, C. & Weiss, Y.) 3–19 (Springer, 2018).
41. Liu, J., Tang, J. & Wu, G. Residual feature distillation network for lightweight image super-resolution. In *Proc. European Conference on Computer Vision (ECCV)* (eds Bartoli, A., Fusiello, A.) 41–45 (Springer, 2021).
42. Alvarez-Ramos, V., Ponomaryov, V. & Sadovnychiy, S. Image super-resolution via wavelet feature extraction and sparse representation. *Radioengineering* **27**, 603 (2018).
43. Xie, C. et al. Large Kernel distillation network for efficient single image super-resolution. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 1283–1292 (IEEE, 2023).

Acknowledgements

This study was supported by the National Key R&D Program of China (Nos. 2023YFF0906700 and 2023YFF0906704) and the National Natural Science Foundation of China (No. 41571369).

Author contributions

Y.C.: Investigation, Conceptualization, data curation, methodology, software, visualization, writing—original draft, writing—review and editing. A.Z.: Conceptualization, methodology, supervision, resources, writing—review and editing. F.G.: Investigation, methodology, project administration, supervision, resources, writing—review and editing. J.G.: Software, data curation, investigation, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Aiwu Zhang or Feng Gao.