Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Visual Context-Aware and Spatial Receptive Field-Based algorithm for mural inpainting

Yang Liu
Lanzhou University of Technology
Zhongmin Liu
liuzhmx@163.com
Lanzhou University of Technology
Wenjin Hu
Northwest Minzu University

Article

Keywords: Dunhuang Mural Restoration, Prompt Vectors, Sliding Window Attention, Spatially Perceptive Feedforward Network

Posted Date: June 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9273578/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Visual Context-Aware and Spatial Receptive Field-Based algorithm for mural inpainting

Yang Liu[1], Zhongmin Liu[1]*, Wenjin Hu[2]

1 College of Electrical Engineering and Information Engineering, Lanzhou University of Technology, Lanzhou, 730050, China

2 College of Mathematics and Computer Science, Northwest Minzu University, Lanzhou, 730030, China

* Corresponding author: College of Electrical Engineering and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China (Zhongmin Liu) E-mail address: liuzhmx@163.com

Abstract

To address the issue of local pixel forgetting and insufficient spatial perception during the restoration of Dunhuang murals, which leads to the distortion of local details in certain areas, we propose a mural restoration algorithm based on visual context awareness and spatial receptive fields. First, a damaged mural is input into an encoder to extract texture and color features, and a residual state-space module is used to model global dependencies. Guidance vectors are used to learn similar semantic pixels, which compensates for missing local details, while local convolutions, along with channel attention mechanisms, restore neighboring details. A multi-scale deformable sliding window attention module is introduced in the decoder to capture multi-level image features, from coarse to fine, enhancing the representation of irregular spatial patterns. Finally, a spatially perceptive feedforward network strengthens long-range dependencies and positional awareness. Experimental results on the Dunhuang dataset demonstrate that the proposed algorithm improves PSNR by 0.18%-0.94%, SSIM by 1.17%-2.98%, reduces L1 by 16.5%-35.92%, and decreases LPIPS by 6.53%-14.31%.

Keywords: Dunhuang Mural Restoration, Prompt Vectors, Sliding Window Attention, Spatially Perceptive Feedforward Network

Introduction

The murals of the Mogao Grottoes, a vast complex of caves constructed between the 4th and 14th centuries, reflect the historical development of cultural exchange along the Silk Road. They serve as invaluable records for the study of the development of ancient Chinese Buddhist art. These murals not only hold profound religious and cultural significance, along with remarkable artistic achievements, but also contain valuable scientific and technological insights, such as ancient pigment preparation techniques and the transmission of painting methods. As such, they possess irreplaceable historical, artistic, and scientific research value.

However, due to the location of the Mogao Grottoes in the extremely arid desert region at the western end of the Hexi Corridor, the murals face multiple degradation threats in this harsh natural environment. These threats primarily involve mural shrinkage and cracking caused by prolonged drought, surface wear due to wind and sand erosion, and salinization resulting from the migration and accumulation of soluble salts in groundwater. The combined effects of these factors have led to typical structural diseases in the murals, such as chalking, blistering, and peeling. The mechanisms underlying these diseases involve intricate physical and chemical processes, directly affecting the adhesion stability of the pigment layers and posing a significant threat to the overall integrity and aesthetic quality of the murals.

In response to the serious conservation challenges faced by the Dunhuang murals, the Dunhuang Academy has established a relatively comprehensive theoretical system for mural conservation after decades of research and practice. They have also developed scientific technical specifications and operational standards. However, traditional physical restoration methods, such as grouting reinforcement, surface cleaning, and pigment repainting, while able to control the progression of degradation to some extent, have inherent limitations, such as irreversible repair processes and high operational risks. Furthermore,

traditional restoration work requires a high level of expertise, has long restoration cycles, and incurs significant economic costs, making it difficult to meet the demands of large-scale mural conservation[1].

In recent years, with the rapid development of computer science, digital image processing technologies, deep learning algorithms, and computer vision theories, new technical solutions have emerged for the field of cultural heritage preservation. Virtual restoration technologies based on digital image processing offer significant advantages, such as non-contact operation, repeatability, and cost-effectiveness. These technologies can reconstruct missing information and restore the visual appearance of murals without any physical intervention to the original cultural relics. This technological approach not only opens new directions for the digital protection, inheritance, and exhibition of the Dunhuang murals but also lays a solid theoretical foundation and technical support for establishing a modern cultural heritage preservation system. The continuous emergence of high-performance methods in the field of image restoration has played a crucial role in the restoration of cultural relic images.

Generative Adversarial Networks (GANs) have demonstrated unique advantages in the field of cultural heritage image restoration. For common issues such as wear, cracks, and fading in cultural relic images, GANs can utilize surrounding contextual information to generate missing content that is highly consistent with the original artwork style, thereby restoring details while maintaining the overall structural integrity. For example, the EdgeConnect model proposed by Nazeri et al. [2] uses an edge generator to predict the structural outlines of damaged areas, followed by an image completion network to fill in color and texture details, achieving significant progress in visual coherence and structural rationality. Lin et al. [3] proposed the PD-GAN model, which generates multiple high-fidelity restoration results for the same damaged region through two adversarial networks optimized cooperatively. Zheng et al. [4] introduced the Transformer into GANs and proposed the multivariate image completion framework VQGAN, which not only maintains image quality and diversity but also achieves faster inference speeds. However, when restoring culturally rich and uniquely styled relic images, GANs often struggle to accurately reproduce historical elements and cultural features. They may introduce noise during the restoration process, leading to discrepancies with the intrinsic logic of the original image, such as incorrect texture direction or shape distortion. These issues limit the application of GANs in high-precision cultural relic restoration.

On the other hand, Convolutional Neural Networks (CNNs) are commonly used as the backbone networks for image restoration due to their powerful ability to capture local patterns and textures. CNNs excel at handling regions with rich details but are often limited by receptive field constraints and inefficient use of information when processing broader image contexts. To address this, researchers have attempted to combine CNNs with large-kernel convolutions. Ding et al. [5] proposed the RepLKNet architecture, which uses large convolution kernels to build larger effective receptive fields, preserving CNN's advantages in local feature extraction while enhancing its ability to model long-range dependencies. However, traditional CNN methods rarely analyze statistical correlations between intermediate layers or channels, and their utilization of low-frequency information is insufficient, limiting their representational capacity and hindering further improvements in reconstruction quality. To overcome these limitations, Dai et al. [6] designed a multi-scale non-local and residual attention framework that expands the effective receptive field without significantly increasing the number of parameters, while balancing high-frequency details and low-frequency information flow. Additionally, by introducing channel attention mechanisms into global covariance pooling, the model can better capture rich correlations between channels and focus on high-frequency features with more information. Hu et al. [7] proposed the Squeeze-and-Excitation module to explicitly model channel dependencies, enhancing the network's feature representation ability and improving restoration accuracy. However, traditional residual dense networks primarily rely on convolution

operations, making it difficult to effectively integrate long-range contextual information, especially when recovering repetitive textures or long-distance structures.

To address this, researchers have introduced Transformers into image restoration to fully leverage their ability to capture global correlations. Song et al. [8] proposed a two-stage contextual image restoration framework, where GAN is first used to roughly complete missing areas, followed by an attention module that learns surrounding contextual information to generate more natural restoration results. Liu et al. [9] introduced the consistent semantic attention method, which analyzes the semantic relationships between hole features to preserve the overall structure of the image and efficiently predict missing content. Li et al. [10] proposed the MAT model, combining the advantages of Transformers and CNNs specifically for restoring large missing areas. In the same year, Wan et al. [11] combined Transformers with CNNs to propose a high-fidelity multivariate restoration method. Additionally, Dong et al. [12] achieved high-quality restoration by efficiently reasoning in the low-resolution sketch space and using an attention mechanism to progressively rebuild the image structure. To further improve detail consistency, Liang et al. [13] proposed a residual Swin Transformer restoration framework that combines Swin Transformer's window-shifting mechanism with residual convolutions, maintaining local computational efficiency while enabling global interaction across windows and reducing blocky artifacts. However, window attention has limitations, such as insufficient information exchange between windows, which can result in "blocky traces" in intermediate features. To address this, Chen et al. [14] introduced channel attention alongside the standard window self-attention mechanism, activating a wider range of pixels through global channel statistics and introducing larger overlapping windows for cross-window interaction, thereby eliminating blocky artifacts and improving overall restoration consistency. These advancements are largely attributed to the continuous expansion of the network's receptive field, which allows the model to gather information from broader regions and use more pixels to reconstruct the target area. It also aids in extracting high-level patterns and structures from images, which is crucial for maintaining image integrity. However, computational efficiency remains a challenge when processing long sequence data.

Although Transformer models have demonstrated excellent performance in image restoration, they typically rely on single-modality input and are prone to overfitting when training data is insufficient. To address this, researchers have started incorporating large language models into prediction tasks to leverage their powerful contextual modeling capabilities. Liu et al. [15] proposed the CALF framework, which fine-tunes and aligns large language models across modalities, reducing the distribution discrepancy between text and time-series data and significantly enhancing multivariate prediction performance.

The Mamba method combines the advantages of Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), achieving near-linear time complexity, making it well-suited for long-sequence data. Gu et al. [16] introduced the S4 sequence model based on state-space models, significantly improving the computational and memory efficiency of SSMs by innovative parameterization and optimization techniques. Later, Gu et al. [17] proposed the Selective State-Space Model, which uses a selective mechanism and hardware-aware algorithm for linear-time sequence modeling, demonstrating superior performance over Transformers across various fields. Hu et al. [18] proposed the Zigzag Mamba method, which rearranges image block sequences to maintain spatial continuity during Mamba's scanning process. Islam et al. [19] introduced the TranS4mer model, which integrates the short-range modeling capability of Transformers with Mamba’s long-range modeling advantages to accurately detect image boundary areas. Qin et al. [20] proposed the Mamba VC Visual Compression Network, which introduces visual state-space blocks and 2D selective scanning modules, enhancing global context modeling and compression efficiency effectively. Xing et al. [21] presented the Seg Mamba image segmentation model, which uses TS

Mamba, GSC, and FUE modules to efficiently model long-range dependencies in image features while maintaining computational efficiency. Chen et al. [22] combined the global interaction capability of Transformers with Mamba for collaborative pixel- and block-level interaction learning. This model employs self-attention mechanisms in the frequency domain to capture long-range dependencies and assigns different attention weights based on the importance of frequency bands, significantly improving image restoration quality and contextual accuracy. However, since the model relies on causal state-space modeling, it can only perceive prior information and cannot utilize subsequent pixels, leading to insufficient information usage. Moreover, long-range interaction attenuation leads to local detail loss, thus limiting performance in fine-grained restoration.

To address the issues of local pixel forgetting and insufficient spatial awareness, we propose a mural restoration algorithm based on visual context awareness and spatial receptive fields (Visual Context-Aware and Spatial Receptive Field-Based Algorithm for Mural Inpainting, VCSR) to improve restoration performance. The main contributions of this paper are as follows:

1. We design a Residual State-Space Block (RSSB), which includes the Attentive State-Space Module (ASSM), local convolutions, and channel attention. The ASSM introduces semantic cues into the initial state-space equations, enabling each pixel query to break free from the traditional scanning order and achieve global awareness. This effectively addresses receptive field limitations and local pixel forgetting. At the end of the RSSB block, local convolutions refine neighborhood features for more accurate detail restoration. These convolutions, while refining features, also incorporate learnable channel attention, which automatically identifies and enhances key channel representations by focusing on important channels and avoiding redundant information.

2. We propose a Multi-Scale Deformable Sliding Window Attention Module (MSDA). Building on traditional multi-scale attention mechanisms, MSDA introduces deformable offsets that dynamically adjust the position and shape of the attention window, expanding the attention range from fixed small windows to broader potential key areas. This improves feature extraction capabilities and enhances information interaction across different scales and regions.

3. We introduce a Spatially Perceptive Feedforward Network (SPFN). The SPFN extracts spatial information at both the input and output ends of the MSDA module and integrates this information into the main feature through a gating mechanism. This enhances spatial dependencies and positional information, improving the reconstruction quality of local details.

Methods

This paper proposes a mural restoration algorithm based on visual context awareness and spatial receptive fields. The architecture follows a U-Net structure, consisting of an encoder and a decoder. The overall network framework is illustrated in Figure 1.

In the encoding phase, the damaged mural, along with the added mask, is fed into the encoder. A 3×3 convolutional layer is first applied to extract initial texture and detail information. A four-stage hierarchical structure is employed for deep feature extraction, where each stage halves the resolution, with skip connections employed to retain fine-grained features. In the RSSB module, the Attentive State-Space Module incorporates global semantic cues to capture long-range dependencies in the mural image, effectively mitigating the loss of neighborhood information and detail distortion during feature flattening. Local convolutions and channel attention further refine key channels and local details.

In the decoding phase, a Multi-scale Deformable Sliding Window Transformer Block (MDSTB) module is introduced as part of the decoder. The core of this module includes multi-scale deformable sliding window attention and a spatially perceptive feedforward network, combined with layer normalization and residual connections. First, the deep features output by the encoder are fed into the decoder. The MSDA module dynamically samples sliding windows at multiple scales, with query indicators

positioned at the center, enabling both global and local information interaction across windows. The SPFN module extracts spatial position information before and after the MSDA module and integrates global semantic and structural information via a gating mechanism, thereby enhancing the spatial consistency and accuracy of mural restoration. Finally, an output convolution generates a high-quality mural restoration result with rich details and coherent structure.

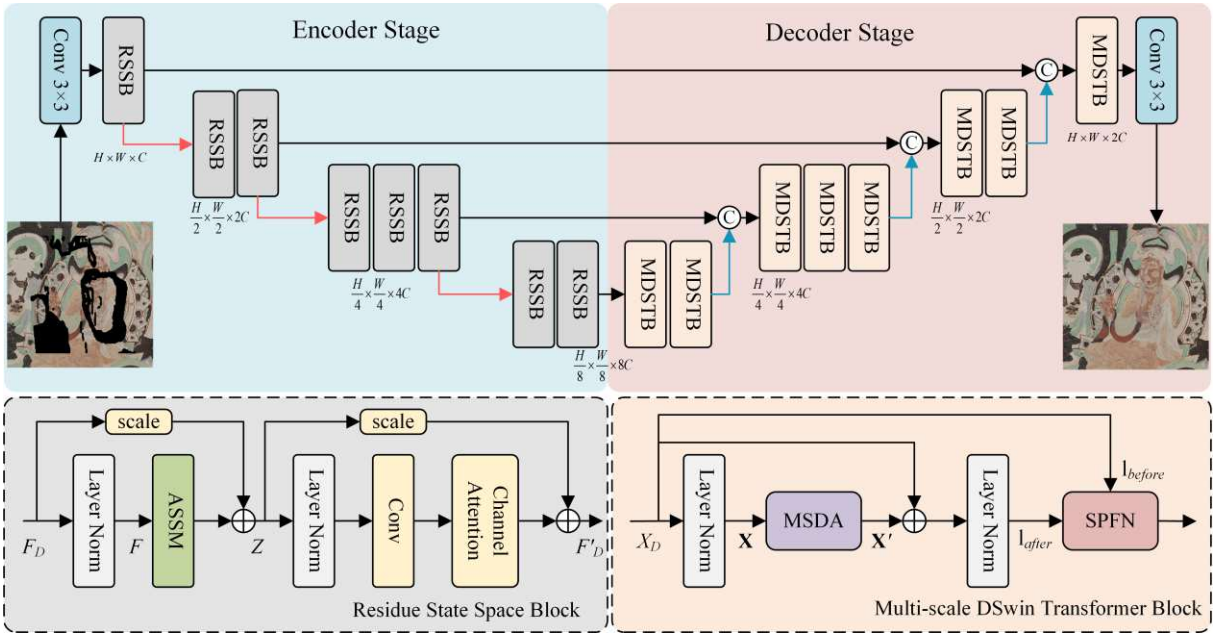


Fig. 1 The architecture of the proposed Network for image inpainting.

Residual State-Space Block.

State-space models typically rely on specific scanning strategies that flatten a 2D image into a 1D sequence for sequential modeling. This increases the distance between adjacent pixels, causing the loss of local pixel information. Additionally, multi-directional scanning increases computational complexity significantly. To address these issues, we propose a Residual State-Space Module. This module aggregates semantically similar pixels via the Attention State-Space Module, compensates for the missing information of unsampled pixels, and restores neighborhood similarity through additional local convolutions, as shown in Figure 1.

Specifically, the given input deep feature F_D is first standardized through Layer Normalization (LN). Then, the Attention State-Space Module efficiently captures the global dependencies of Mamba, and the learnable scaling factor controls the information flow in the skip connections. Finally, the output feature Z is obtained, as shown in Equation (1):

$$Z = \mathrm{ASSM}(\mathrm{LN}(F_D)) + s \cdot F_D \tag{1}$$

In addition, since the State-Space Model (SSM) flattens the 2D feature map into a 1D sequence for processing, adjacent pixels in the spatial domain are placed far apart in the 1D sequence. Furthermore, the four-way scanning strategy can only anchor pixels to the four nearest neighbors. Therefore, the feature Z is first normalized, and an additional local convolutional layer is added after the Attention State-Space Module (ASSM) to compensate for the loss of local features. Additionally, SSM typically requires a large number of hidden states to capture long-range dependencies, which may lead to channel redundancy. To enhance the expressive power of different channels, the Residual State-Space Block (RSSB) module introduces a channel attention mechanism, enabling the SSM to learn more diverse channel representations. This allows subsequent channel attention to select key channels, thereby avoiding channel redundancy. Finally, an adjustable scaling factor is applied to the residual connection to obtain the final output feature F'_D of the RSSB, as shown in Equation (2):

$$F'_D = \mathrm{CA}(\mathrm{Conv}(\mathrm{LN}(Z))) + s' \cdot Z \tag{2}$$

Comparison between Attention Mechanisms and

State-Space Models.

Existing state-space model methods primarily rely on the Mamba architecture for image reconstruction, where discrete state-space equations are used to model the interactions between pixel blocks:

$$\begin{aligned} h_i &= \bar{\mathbf{A}} h_{i-1} + \bar{\mathbf{B}} x_i \\ y_i &= \mathbf{C} h_i + \mathbf{D} x_i \end{aligned} \tag{3}$$

where $\bar{\mathbf{A}}$ =exp(Δ**A**) is the control matrix, $\bar{\mathbf{B}}$ =(Δ**A**)$^{-1}$(exp(Δ**A**)-I)(Δ**B**)≈Δ**B** is the input matrix, and **C** is the output matrix. Equation (3) shows that under this strategy, the *i*-th pixel can only observe a limited number of *i*-1 pixels from the entire image, which prevents the full utilization of globally similar pixel information. Additionally, there is the issue of long-range information decay, which leads to the loss of crucial pixel information.

To introduce non-causal modeling capabilities into Mamba, this paper unifies the causal linear attention mechanism with the state-space model into a common framework and analyzes the relationship between the two. First, given the query matrix Q, key matrix K, and value matrix V, the output of the linear attention mechanism is computed as shown in Equation (4):

$$y_i = \sum_{j=1}^{i} \frac{\mathbf{Q}_i \mathbf{K}_j^{\cdot}}{\sum_{t=1}^{i} \mathbf{Q}_i \mathbf{K}_t^{\cdot}} \mathbf{V}_j = \frac{\mathbf{Q}_i \left(\sum_{j=1}^{i} \mathbf{K}_j^{\cdot} \mathbf{V}_j \right)}{\mathbf{Q}_i \left(\sum_{t=1}^{i} \mathbf{K}_t^{\cdot} \right)} \tag{4}$$

where $\mathbf{S}_i = \sum_{j=1}^{i} \mathbf{K}_j^{\cdot} \mathbf{V}_j$, $\mathbf{Z}_i = \sum_{t=1}^{i} \mathbf{K}_t^{\cdot}$, The formula for linear attention is given by:

$$y_i = \frac{\mathbf{Q}_i \mathbf{S}_i}{\mathbf{Q}_i \mathbf{Z}_i} \tag{5}$$

where $\mathbf{S}_i$=$\mathbf{S}_{i-1}$+ $\mathbf{K}_i^{\cdot}$ $\mathbf{V}_i$, $\mathbf{Z}_i$=$\mathbf{Z}_{i-1}$+ $\mathbf{K}_i^{\cdot}$. For ease of observing the relationship, the formula is further restructured into a unified form as follows:

$$\begin{aligned} \mathbf{S}_i &= \mathbf{I}\mathbf{S}_{i-1} + \mathbf{K}_i^{\cdot} \mathbf{V}_i, \\ y_i &= \mathbf{Q}_i \mathbf{S}_i / \mathbf{Q}_i \mathbf{Z}_i + \mathbf{O} x_i, \end{aligned} \tag{6}$$

where **I** and **O** represent the identity matrix and the zero matrix, respectively, and x_i denote the input label at the *i*-th step. In the state-space equation (3), $\bar{\mathbf{B}}$ x_i ≈ Δ**B**(x_i)=**B**(Δx_i). Thus, the state-space equation can be reconstructed into a unified form as follows:

$$\begin{aligned} h_i &= \bar{\mathbf{A}} h_{i-1} + \bar{\mathbf{B}} (\Delta x_i) \\ y_i &= \mathbf{C} h_i / \mathbf{I} + \mathbf{D} x_i \end{aligned} \tag{7}$$

By comparing Equations (6) and (7), it can be seen that there is a close relationship between the attention mechanism and the state space, i.e., h_i ~ $\mathbf{S}_i$、**B** ~ $\mathbf{K}^{\mathrm{T}}$、**C** ~ **Q**. The output matrix **C** in Equation (3) serves a similar role to the query in the attention mechanism. The core idea is to integrate the information of the unsampled sequence into **C**, allowing **C** to "query" the unsampled pixels in the form of attention, thereby assisting in the restoration of the input x_i.

Attentive State-Space Module.

To efficiently leverage Mamba for capturing global dependencies while further considering the modeling of local interactions, an Attention State-Space Module is proposed, as shown in Figure 2. For a given input feature $F \in \mathrm{R}^{H \times W \times C}$, where *H* and *W* represent the height and width of the feature map, and *C* denotes the number of channels, position encoding is first applied to *F* to preserve the original structural information. Then, a Semantic Guided Neighboring (SGN) strategy is introduced, which unfolds the 2D image into a 1D sequence for subsequent modeling using the Attentive State-Space Equation (ASE). Finally, another SGN is used as the inverse operation of the unfolding process to reassemble the sequence back into the image and obtain the output feature of the module through linear mapping.

Specifically, the Attention State-Space Equation aims to modify the output matrix $\mathbf{C} \in ¡^{L \times d}$, where *L*=*H*×*W* is the length of the sequence after the image is flattened, and *d* represents the hidden state dimension in the Mamba model. To query relevant pixels in the image globally, the Attention State-Space Equation is proposed, incorporating the prompt vectors into the matrix **C.** These prompt vectors can learn and represent a set of pixels with similar semantics, thereby supplementing the missing information for unobserved pixels. First, a prompt pool $\mathrm{P} \in \mathrm{R}^{T \times d}$ is constructed, where *T* is the number of prompts in the pool. To enhance interpretability, semantic decoupling parameterization is applied to the

prompt pool P ：

$$\mathrm{P} = \mathbf{MN},\ \mathbf{M} \in ¡^{T\times r},\ \mathbf{N} \in ¡^{r\times d} \quad (8)$$

where **N** represents the shared similarity feature space between different modules, **M** is the feature combination coefficient for each module, and r is the inner rank, satisfying $r \ll \min\{T,d\}$. The semantic decoupling not only ensures consistency across modules but also allows for local differentiated modeling, thereby enhancing the overall network's expressive power and interpretability.

Next, a selection mechanism is designed to choose L specific prompt instances from the prompt pool P based on the input features, forming matrix $\mathbf{P} \in \mathrm{R}^{L\times d}$, which is then added to the output matrix **C** to incorporate information about the unobserved pixels. Specifically, given the flattened input feature $\mathrm{F}' \in \mathrm{R}^{L\times C}$, a linear transformation is first applied to project its channel dimension from C to T. Then, the LogSoftmax function is used to compute the log probabilities, measuring the likelihood that each prompt in the prompt pool is sampled by the i-th input label $F_i(i=1,2,\cdots, L)$. Subsequently, Gumbel-Softmax is applied to the log probabilities, enabling differentiable prompt selection and resulting in a selection matrix $\mathbf{R} \in \mathrm{R}^{L\times T}$. Through matrix multiplication **P**=**R** P, the corresponding instance-specific prompt matrix $\mathbf{P} \in \mathrm{R}^{L\times d}$ is generated. Finally, residual addition is used to integrate **P** into **C**, constructing the final Attention State-Space Equation as follows:

$$\begin{aligned} h_i &= \bar{\mathbf{A}} h_{i-1} + \bar{\mathbf{B}} x_i \\ y_i &= (\mathbf{C} + \mathbf{P}) h_i + \mathbf{D} x_i \end{aligned} \quad (9)$$

The learned prompts enable the model to acquire an attention mechanism-like capability, allowing it to query relevant pixels across the entire image. By introducing prompts that represent sets of similar pixels, the proposed ASE effectively alleviates the issue of limited perception of unobserved pixels. Furthermore, this method requires only a single-direction scan for modeling, thereby avoiding the high computational cost and redundancy issues associated with multi-directional scanning in existing methods.

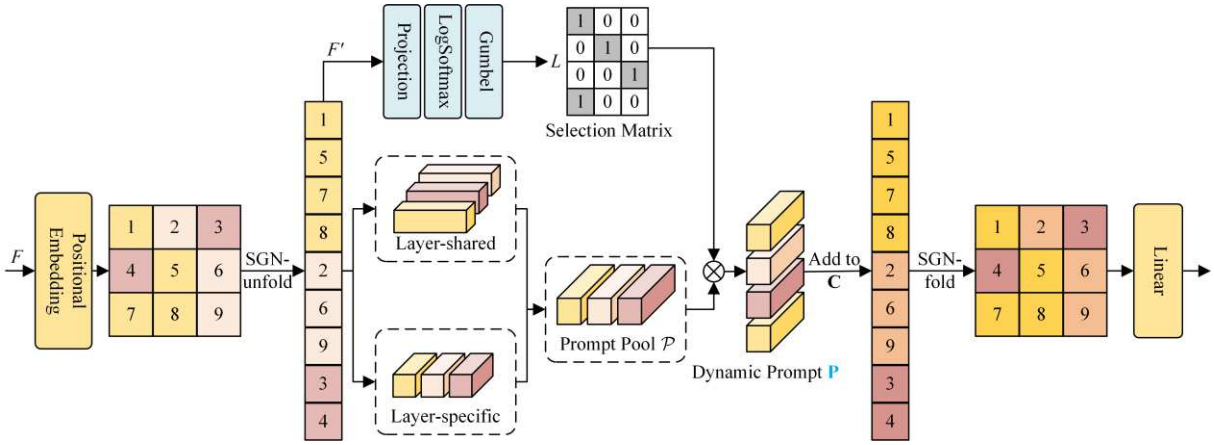


Fig. 2 ASSM structure framework

Multi-Scale Deformable Sliding Window Attention Module.

To address the limitations of traditional window-based self-attention mechanisms in image modeling, such as restricted cross-window interaction and limited receptive field range, a Multi-Scale Deformable Sliding Window Attention (MSDA) module is proposed, as shown in Figure 3. By combining deformable sampling with multi-scale processing, the MSDA module enables adaptive aggregation of features across different spatial ranges, allowing the attention to focus more on key regions and effectively capture both local and global contextual information in the image.

Specifically, the input feature map is denoted as $\mathbf{X} \in \mathrm{R}^{H\times W\times C}$. The self-attention mechanism calculates the attention weights by comparing each query feature $\mathbf{x}_{i,j}$ with other features within its receptive field. To incorporate local contextual information, the attention weight at position (i,j) is defined as shown in Equation (13):

$$\alpha_{i,j}^{(u,v)} = \frac{\exp\left(\mathbf{x}_{i,j}, \mathbf{x}_{i+u,j+v}\right)}{\sum_{(u',v')\in \mathrm{N}_k} \exp\left(\mathbf{x}_{i,j}, \mathbf{x}_{i+u',j+v'}\right)} \quad (10)$$

where $(u,v) \in N_k$ represents the local neighborhood defined by the kernel size k . The output feature at this position $\mathbf{Y}_{i,j}$ is computed as follows:

$$\mathbf{Y}_{i,j} = \sum_{(u,v)\in \mathrm{N}_k} \alpha_{i,j}^{(u,v)} \mathbf{x}_{i+u,j+v} \quad (11)$$

This formula limits the attention computation to the local neighborhood, similar to the convolution operation, which not only improves computational efficiency but also emphasizes the critical role of the receptive field in this mechanism. To adaptively expand the receptive field, deformable offsets are introduced into the attention mechanism. Specifically, for each position (i,j) and each position (u,v) within its local neighborhood, an offset is learned as follows:

$$\Delta \mathbf{p}_{i,j}^{(u,v)} = f_\theta(\mathbf{x}_{i,j}) \quad (12)$$

where f_θ is a lightweight module used to predict the offset for the sampling positions.

Where $\Delta u_{i,j}^{(u,v)}$ and $\Delta v_{i,j}^{(u,v)}$ represent the horizontal and vertical components of the offset vector $\Delta \mathbf{p}_{i,j}^{(u,v)}$, respectively. Finally, the output feature is obtained by merging the sampled features with the adaptively selected attention weights, as shown below:

$$\mathbf{Y}_{i,j} = \sum_{(u,v)\in \mathrm{N}_k} \alpha_{i,j}^{(u,v)} \mathbf{x}_{i+u+\Delta u_{i,j}^{(u,v)}, j+v+\Delta v_{i,j}^{(u,v)}} \quad (13)$$

The basic sliding window attention mechanism is further extended into a multi-scale variant, called the Multi-Scale Deformable Sliding Window Attention Module. It adjusts the receptive field by assigning different kernel sizes to different attention heads, capturing rich hierarchical features. Formally, there be H attention heads, where each attention head $h \in \{1,\cdots,H\}$ corresponds to a unique kernel size k_h. The attention computation process for the h-th attention head is as follows:

$$\mathbf{Y}_{i,j}^{h} = \sum_{(u,v)\in \mathrm{N}_{k_h}} \alpha_{i,j,h}^{(u,v)} \mathbf{x}_{i+u+\Delta u_{i,j}^{h}, j+v+\Delta v_{i,j}^{h}} \quad (14)$$

where N_{k_h} represents the definition of the kernel size k_h used by the h-th head, and $(\Delta u_{i,j}^{h}, \Delta v_{i,j}^{h})$ is the deformable offset learned for the h-th head. Finally, the output features from different attention heads are concatenated through a linear projection:

$$\mathbf{Y}_{i,j} = \mathbf{W}_o[\mathbf{Y}_{i,j}^{1}; \mathbf{Y}_{i,j}^{2}; \ldots; \mathbf{Y}_{i,j}^{H}] \quad (15)$$

where $\mathbf{W}_o \in \mathrm{R}^{C\times C}$ represents the output projection matrix, and the symbol [;] indicates the concatenation of features.

By introducing deformable offsets, the receptive field can be adaptively adjusted, allowing the attention mechanism to focus on relevant regions beyond the fixed local window. This enhances the model's ability to model irregular spatial relationships. As a result, the model can flexibly capture the potential dependencies between different regions in the image, improving its ability to recognize complex structures and details.

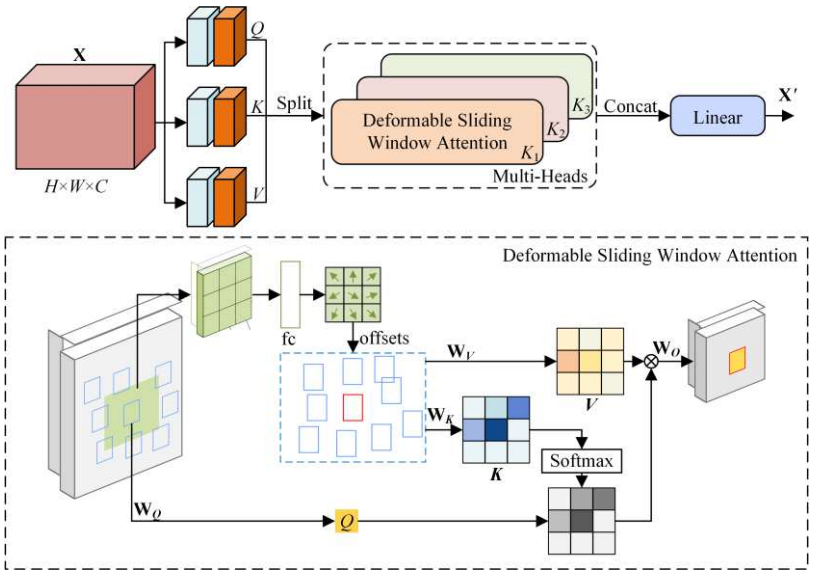


Fig. 3 MSDA structure framework.

Spatially Perceptive Feedforward Network. To address the inherent limitations of the State-Space Model (SSM) in handling local spatial information that spans multiple rows and columns, a Spatially Perceptive Feedforward Network (SPFN) is proposed to enhance spatial awareness in image representation learning. As shown in Figure 4, SPFN extracts spatial information from the feature representations before entering the MSDA module. This information is then utilized through a gating mechanism to guide the MSDA module in processing subsequent features, thereby promoting the deep integration of spatial awareness and long-range dependency learning throughout the MSDA module.

Specifically, SPFN first extracts the features $\mathbf{I}_{before}$ and $\mathbf{I}_{after}$ from the input and output of the MSDA module, respectively. Next, SPFN applies an average pooling operation to $\mathbf{I}_{before}$ to expand the receptive field and uses a two-layer Conv-LN-ReLU combination module to capture broader spatial information. Subsequently, an upsampling operation is performed to generate a spatially perceptive indicator γ that retains the original spatial relationships. In the gating mechanism, $\mathbf{I}_{after}$ is divided into $\mathbf{I}'_{after}$ and $\mathbf{I}''_{after}$. Guided by the spatial perceptive indicator γ, $\mathbf{I}'_{after}$ undergoes a linear transformation and *GELU* non-linearity to generate the gating signal, which modulates $\mathbf{I}''_{after}$ through element-wise multiplication, significantly enhancing its spatial awareness. The overall process is defined as follows:

$$\mathbf{I}'_{after} = W_{d3}W_1\mathrm{LN}(\mathbf{I}_{after}) \quad (16)$$

$$\mathbf{I}''_{after} = W'_{d3}W'_1\mathrm{LN}(\mathbf{I}_{after}) \quad (17)$$

$$\gamma = Up(f(AveragePooling(\mathbf{I}_{before}))) \quad (18)$$

$$gate = \mathrm{GELU}(W_{d3}W_1\gamma \,\|\, \mathbf{I}'_{after}) \quad (19)$$

$$output = gate \,\mathrm{e}\, \mathbf{I}''_{after} \quad (20)$$

where W_1 and W'_1 represent 1×1 convolution operations, while W_{d3} and W'_{d3} are 3×3epthwise separable convolutions, which refine the features while reducing computational costs. *LN* stands for Layer Normalization, f denotes the combination of two layers of *Conv-LN-ReLU* structure, *Up* represents the upsampling operation, and ⊙ indicates element-wise multiplication.

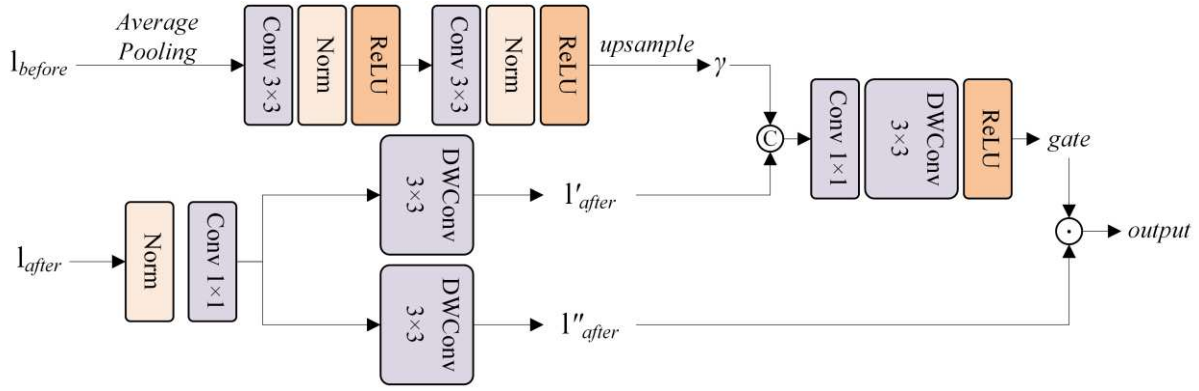


Fig. 4 SPFN structure framework

Experiments

Dataset Settings.

To validate the effectiveness of the proposed image inpainting method, comparative experiments will be performed on the Dunhuang mural dataset, with detailed analysis of the results. Furthermore, to investigate the impact of the designed modules on restoration performance, corresponding ablation studies will be conducted.

The Dunhuang mural dataset used in this study is derived from the Dunhuang Mural Challenge dataset, provided by the Dunhuang Academy. This dataset contains 2,850 mural images with a resolution of 256× 256. Among them, 2,700 images are used for training, 50 images for validation, and 100 images for testing, covering a wide range of subjects, including flying apsaras, arhats, temples, and costumes.

To evaluate restoration performance under varying degrees of damage, the masked images are categorized into five groups based on the proportion of the damaged area: 1%-10%, 10%-20%, 20%-30%, 30%-40%, and random powder masks. The size of the masked region directly correlates with the severity of the image damage. For each category, training data are

generated using various masking strategies, and these data are used to train the model. In addition, 19 authentically damaged Dunhuang mural images are selected to verify the algorithm's restoration effectiveness in real-world scenarios.

Experimental Platform and Evaluation Indicators. The experimental setup is as follows: an 11th Gen Intel(R) Core(TM) i7-11700K @ 3.60GHz processor, with Ubuntu 22.04.5 LTS, and an NVIDIA GeForce RTX 3080 Ti GPU. All experiments were conducted under the same hardware and software configuration.

The environment was built on Python 3.8, and the training process was implemented using the PyTorch deep learning framework. The training consisted of 600,001 steps, corresponding to 445 epochs. The initial learning rate was 0.0001, with a 0.1 ratio between the discriminator's and generator's learning rates.

Simulated Inpainting of Damaged Murals: Experiment and Analysis. To validate the effectiveness of the VCSR algorithm, this experiment simulated mural damages of varying severity. In the test results, five Dunhuang mural images were randomly selected and subjected to restoration under both 1%-40% masking and powder masks. Subsequently, the restoration results of the VCSR algorithm were compared with those of MISF [23], PUT [24], HINT [25], ECDB [26], SEM [27], and MXT [22], through both objective metrics and subjective visual analysis.

Quantitative Analysis. To ensure the comprehensiveness and objectivity of the restoration performance evaluation, four metrics were employed in this experiment: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), and the L1 Mean Error. Specifically, higher PSNR and SSIM values indicate better restoration quality, while lower L1 and LPIPS values also reflect superior restoration performance. The detailed quantitative evaluation results of each algorithm are presented in Table 1:

Table 1 Quantitative comparison of various repair algorithms under different mask rates

Mask rate/contrast algorithm \ Evaluation Indicators		PSNR↑	SSIM↑	L1↓	LPIPS↓
1%～10%	MISF	32.5862	0.9046	0.0106	0.0402
	PUT	32.6916	0.9209	0.0075	0.0367
	HINT	32.7395	0.9186	0.0068	0.0376
	ECDB	32.8933	0.9371	0.0069	0.0341
	SEM	32.9167	0.9446	0.0057	0.0355
	MXT	33.2108	0.9465	0.0055	0.0338
	Ours	**33.3014**	**0.9572**	**0.0045**	**0.0231**
10%～20%	MISF	28.2527	0.8753	0.0199	0.0794
	PUT	28.3582	0.8858	0.0125	0.0713
	HINT	28.3753	0.8935	0.0120	0.0694
	ECDB	28.4792	0.8965	0.0119	0.0682
	SEM	28.6229	0.9001	0.0105	0.0661
	MXT	28.5467	0.8946	0.0103	0.0643
	Ours	**28.8163**	**0.9213**	**0.0066**	**0.0551**
20%～30%	MISF	25.1047	0.8135	0.0217	0.1594
	PUT	25.5767	0.8247	**0.0183**	0.1432
	HINT	25.0749	0.8116	0.0204	0.1387
	ECDB	25.6143	0.8159	0.0196	0.1314
	SEM	25.6799	0.8275	0.0188	0.1261
	MXT	25.9789	0.8313	0.0183	0.1223

	Ours	**26.3823**	**0.8529**	0.0161	**0.1144**
30%~40%	MISF	22.3602	0.6388	0.0337	0.2042
	PUT	22.5143	0.6914	0.0297	0.1936
	HINT	23.5277	0.7189	0.0283	0.1870
	ECDB	23.7352	0.7592	0.0304	0.1783
	SEM	23.5815	0.7519	0.0272	0.1791
	MXT	23.6183	0.7572	0.0281	0.1757
	Ours	**23.8459**	**0.7831**	**0.0252**	**0.1712**
Powder Mask	MISF	32.8061	0.8929	0.0101	0.0395
	PUT	32.9136	0.9277	0.0061	0.0376
	HINT	32.7922	0.9416	0.0055	0.0366
	ECDB	32.9624	0.9382	0.0065	0.0389
	SEM	33.3554	0.9518	0.0062	0.0362
	MXT	33.3875	0.9541	0.0051	0.0341
	Ours	**33.4251**	**0.9662**	**0.0036**	**0.0332**

From Table 1, under all mask ratios, the proposed VCSR algorithm consistently outperforms competing methods in PSNR, SSIM, L1, and LPIPS. VCSR achieves significantly lower LPIPS values than other algorithms across all mask ratios, indicating substantial improvement in perceptual quality. Even under large-scale masking (30%-40%), the PSNR and SSIM values remain superior to those of baseline models, demonstrating the robustness of VCSR in preserving structural information and spatial consistency. Furthermore, in the powder mask test, VCSR's LPIPS value decreases further to 0.0332, while SSIM reaches 0.9662, demonstrating its advantages in detail restoration and stylistic coherence. These results highlight the excellent generalization capability and perceptual restoration performance of VCSR across varying damage severities and mask types.

The superiority of VCSR in objective metrics such as PSNR, SSIM, and restoration accuracy further confirms its high restoration quality. Higher PSNR values reflect the algorithm's outstanding ability to control pixel-level errors, enabling more precise recovery of original grayscale and color information. Higher SSIM values indicate that the algorithm better preserves structural information, which is crucial for image interpretability and semantic consistency. Moreover, improved restoration accuracy directly reflects the algorithm's precision in reconstructing damaged pixels, especially in areas with complex textures and patterns, thereby reducing the risk of erroneous restoration and enhancing reliability.

Qualitative Analysis. To provide a more intuitive verification of the model's restoration performance, experiments were conducted on murals with different mask ratios. Five Dunhuang murals were selected for comparison against competing algorithms, with differences in restoration performance highlighted using blue bounding boxes, as shown in Figure 5. The visualization results present five randomly selected murals from the 100-image test set under various mask ratios. Specifically, Figure 5(a) shows the original murals, Figure 5(b) displays the masked murals, and Figures 5(c)-5(i) illustrate the restoration results of MISF, PUT, HINT, ECDB, SEM, MXT, and the full VCSR algorithm, respectively. The blue bounding boxes clearly indicate the restoration differences across methods.

From Figure 5, it can be observed that the MISF algorithm, which employs multi-level interactive twin filtering by combining image-level and semantic-level filtering, restores much of the semantic and structural information. However, local textures remain blurred or exhibit minor artifacts. For example, under a 10% mask ratio, the flame textures behind the Buddha in Figure 5(c), row 1, are noticeably faded, and original pixels are lost. At 20% masking, facial details are almost

completely absent, while at 30%-40% masking, severe blurring and feature loss occur. The PUT algorithm, which leverages a pyramid VAE architecture with cross-layer latent variable transfer, integrates high-level semantic information with low-level details. However, when the semantic gap between the missing and surrounding regions is large, the results appear blurry or unnatural. For instance, in small masked regions in Figure 5(d), extraneous pixel artifacts are visible, while at 20%-40% masking, restoration traces and structural distortions become prominent.

The HINT algorithm incorporates a mask-aware pixel shuffle downsampling module and a spatial activation channel attention layer. Although it alleviates structural distortion to some extent, it still suffers from local pixel forgetting. At 10%-20% masking, misalignment occurs in the Buddha's eyes and facial contours, and with increasing mask ratios, missing information exacerbates inconsistency between restored structures and contextual surroundings. The ECDB algorithm, based on an enhanced diffusion bridge model, effectively extracts conditional features from low-quality images, focusing on color and edge details. Between 10%-30% masking, ECDB preserves more texture and structural details; however, at 40% masking, structural confusion and distortion appear. The SEM algorithm, through pixel-level modeling and spatial enhancement, captures long-range dependencies and improves spatial perception. Although it restores most global features under large masks, local details are still missing at 10%-30% masking, as shown in Figure 5(g).

The MXT algorithm, which integrates the strengths of Mamba and Transformer, enhances detail reconstruction and spatial coherence. However, due to insufficient utilization of pixel information and inaccurate long-range information transfer, local detail distortions occur in restoration. For example, under 30%-40% masking, details of earrings and beads in Figure 5(h) are missing.

In contrast, the proposed VCSR algorithm, by integrating residual state-space modeling, multi-scale deformable sliding window attention, and the spatially perceptive feedforward network, effectively models global dependencies, reinforces long-range and positional awareness, and generates detailed yet structurally coherent restorations in local regions. As illustrated in Figure 5(i), VCSR produces visually superior results. When processing images with complex textures and edge details, it generates more natural and realistic restorations, with smoother transitions to undamaged regions.

In summary, the proposed VCSR algorithm demonstrates clear superiority over competing methods in terms of visual quality, objective evaluation metrics, and restoration accuracy, highlighting its strong potential for wide application in mural restoration.

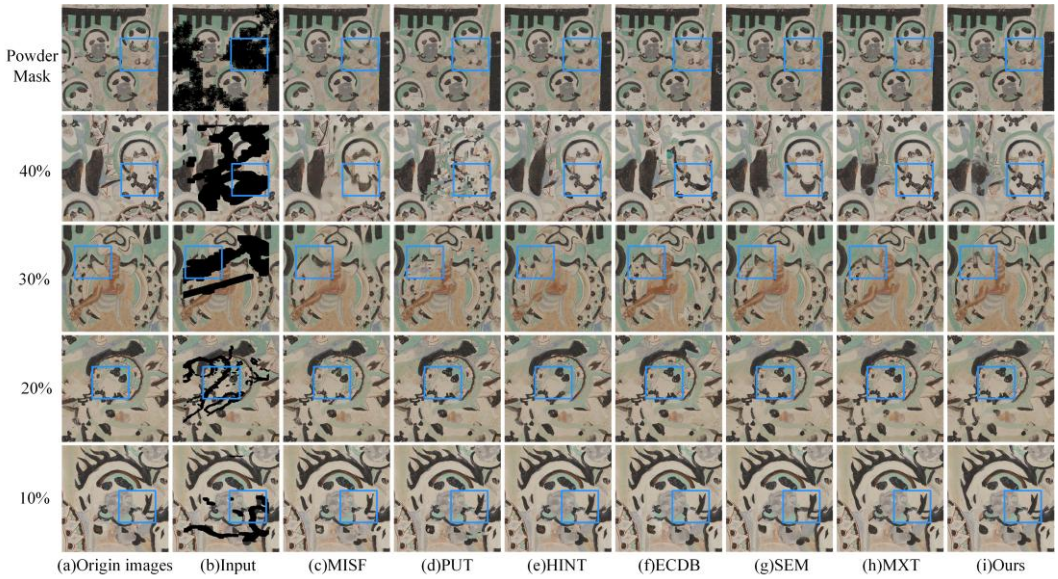


Fig. 5 Comparison chart of the restoration results of each algorithm under different mask rates

Difference Map Analysis. To further validate the effectiveness of the VCSR algorithm, a difference map comparison experiment was conducted under a mask coverage of 10%-20%, with corresponding L1 loss values annotated to provide an intuitive reflection of restoration quality. The visualization results are shown in Figure 6, where Figure 6(a) represents the masked mural and Figure 6(b) shows the original mural. Figures 6(c)-6(p) illustrate the difference maps of MISF, PUT, HINT, ECDB, SEM, MXT, and the proposed VCSR algorithm. In the difference maps, regions closer to zero indicate better restoration quality and smaller errors.

For further analysis, representative murals were selected from the restoration results of all algorithms for detailed comparison. The comparison reveals that the baseline algorithms exhibit significant differences between the restored and original images, highlighting their limitations in detail reconstruction and texture consistency. In contrast, the proposed VCSR algorithm demonstrates smaller residual regions and lower brightness levels in the difference maps, showing a high degree of consistency with the original image. This indicates that VCSR achieves superior visual fidelity in the restoration process.

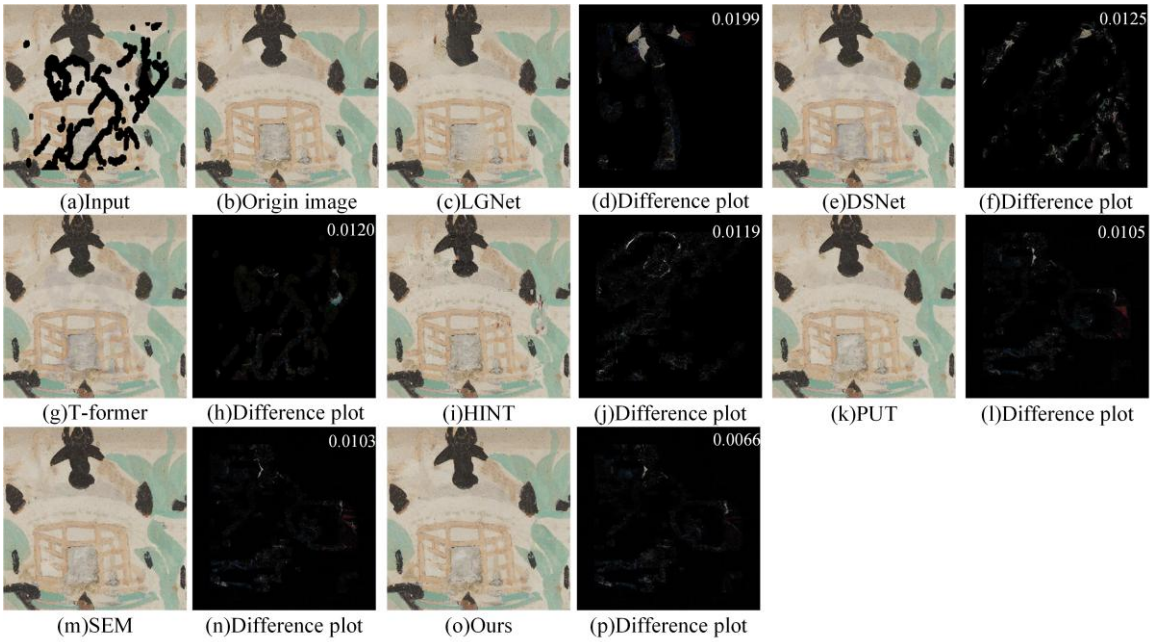


Fig. 6 Comparison algorithm repair difference map

Ablation Experiment and Universality Analysis

Module Ablation Experiment

To evaluate the impact of the proposed modules on restoration performance, experiments were conducted using images with a 10%-20% mask coverage ratio. The RSSB, MSDA, and SPFN modules were incorporated into the baseline model individually and in combination, followed by both quantitative and qualitative analyses.

Quantitative Analysis: The results are presented in Table 2. The second row reports the evaluation metrics of the baseline model, while the third to the sixth rows show the metrics obtained when the RSSB, MSDA, and SPFN modules were introduced individually. The seventh row summarizes the performance when all three modules were integrated simultaneously. The results indicate that incorporating each module independently improves the evaluation metrics, confirming their significant contribution to enhancing restoration quality. Moreover, when all modules are jointly integrated, the image restoration performance is fully optimized, yielding the best overall restoration results.

Table 2 Comparison of quantitative results of module ablation experiments

RSSB	MSDA	SPFN	PSNR↑	SSIM↑	L1↓	LPIPS↓	Params ↓	Elapsed time ↓	GFLOPs ↓

×	×	×	28.5467	0.8946	0.0103	0.0643	38M	22ms	79.16G
✓	×	×	28.6003	0.9051	0.0086	0.0601	32M	20ms	75.25G
✓	✓	×	28.6364	0.9132	0.0075	0.0583	30M	17ms	76.37G
✓	×	✓	28.7207	0.9118	0.0077	0.0597	24M	15ms	76.19G
×	✓	✓	28.7115	0.9193	0.0073	0.0572	25M	14ms	75.42G
✓	✓	✓	**28.8163**	**0.9213**	**0.0066**	**0.0551**	**21M**	**11ms**	**75.15** G

Qualitative Analysis: To further evaluate the synergy of the modules in enhancing mural restoration performance, the visualization results are presented in Figure 7. Figure 7(a) shows the ground truth mural, Figure 7(b) the masked mural, and Figure 7(c) the restoration result of the baseline MXT model. Figures 7(d)-7(h) display the restoration results when incorporating the RSSB, MSDA, RSSB+SPFN, MSDA+SPFN, and the full proposed method, respectively. Areas highlighting restoration differences are marked with blue bounding boxes.

Figure 7(c) shows that the baseline model struggles with large-scale damaged regions, resulting in structural inconsistencies and local detail distortions. The introduction of the RSSB module (Figure 7(d)) notably alleviates edge and local feature loss, primarily due to RSSB's ability to integrate prompt vectors to learn semantically similar pixels and recover local details. However, structural inconsistencies remain, such as misalignment of radiant textures with original line structures in the second mural.

Integrating the MSDA module (Figure 7(e)) further improves the model's spatial information exchange, reducing contextual structural inconsistencies. The inclusion of the SPFN module strengthens the model's ability to perceive and preserve features within damaged areas, enhancing spatial dependencies and positional awareness. In Figure 7(f), the combination of RSSB and SPFN alleviates local texture loss, improving overall authenticity. Figure 7(g) demonstrates that integrating MSDA with SPFN further reinforces structural consistency and ensures effective preservation of high-frequency details during restoration.

Finally, when all modules are combined, the VCSR algorithm achieves significant improvements in clarity and consistency in the restoration results. As shown in Figure 7(h), the restored images exhibit a high degree of visual fidelity, closely approximating the original murals.

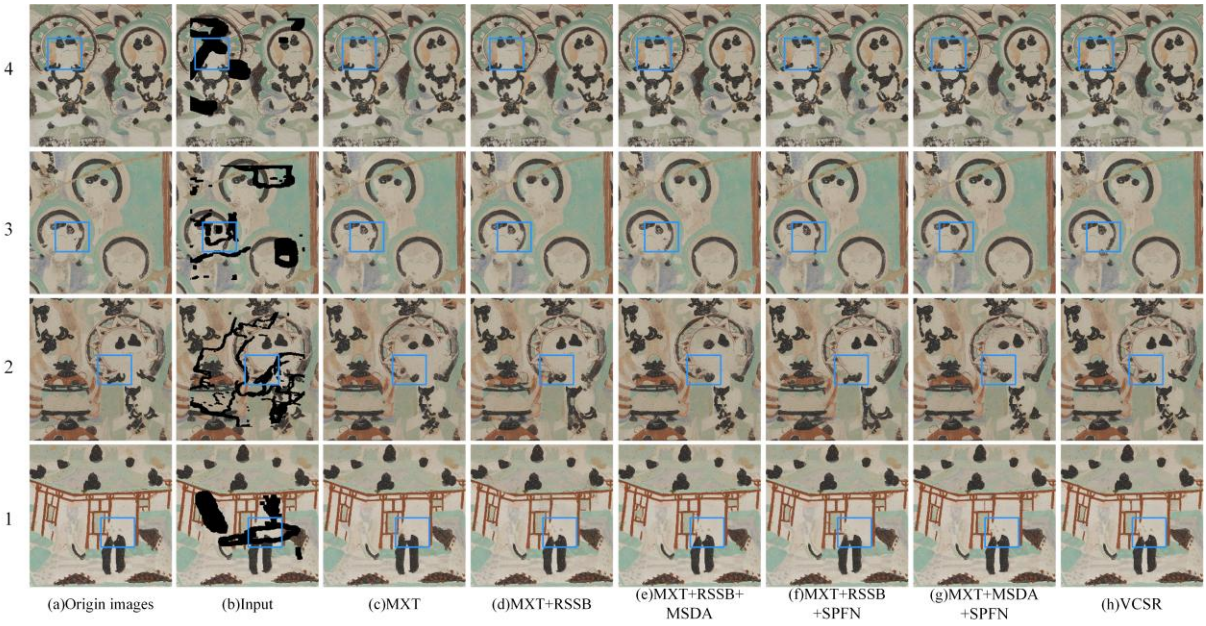


Fig.7 Comparison chart of ablation experiment results

Real damaged mural inpainting experiment. To further verify the effectiveness of the VCSR algorithm in restoring authentically damaged mural images, four real-world damaged murals were randomly selected. Restoration experiments were conducted by applying masks to fill in the damaged regions. As shown in Figure 8, Figure 8(a) represents the original murals, Figure 8(b) shows the masked murals, and Figures 8(c)-8(i) display the restoration results of MISF, PUT, HINT, ECDB, SEM, MXT, and the proposed VCSR algorithm, respectively. Differences in restoration performance across methods are highlighted using blue bounding boxes.

From Figure 8, it can be observed that in the first mural, MISF, PUT, and HINT produce visible blurring in the restored regions. Although ECDB, SEM, and MXT alleviate this issue, they still fail to preserve local details adequately. In contrast, VCSR preserves fine-grained details, avoids structural misalignment, and generates clearer textures with coherent structures—particularly prominent in Figure 8(i), row 1. In the second mural, MISF, PUT, HINT, ECDB, SEM, and MXT exhibit structural misalignments and block-like artifacts in the backlight regions of the Buddha statue, while VCSR successfully retains texture details and structural continuity. In the third mural, all comparison algorithms suffer from unnatural color transitions and structural mismatches at restoration boundaries, whereas VCSR provides more refined and accurate reconstruction of critical features. In the fourth mural, MISF and PUT show large-scale blurring, while the other methods achieve relatively higher-quality restoration.

In summary, the proposed VCSR algorithm demonstrates remarkable advantages in restoring real damaged murals. Its results not only better preserve edge features and global spatial structures, but also achieve superior structural coherence and natural transitions compared with competing methods. Moreover, the restored murals appear visually more vivid and realistic, underscoring the potential of VCSR in practical cultural heritage restoration.

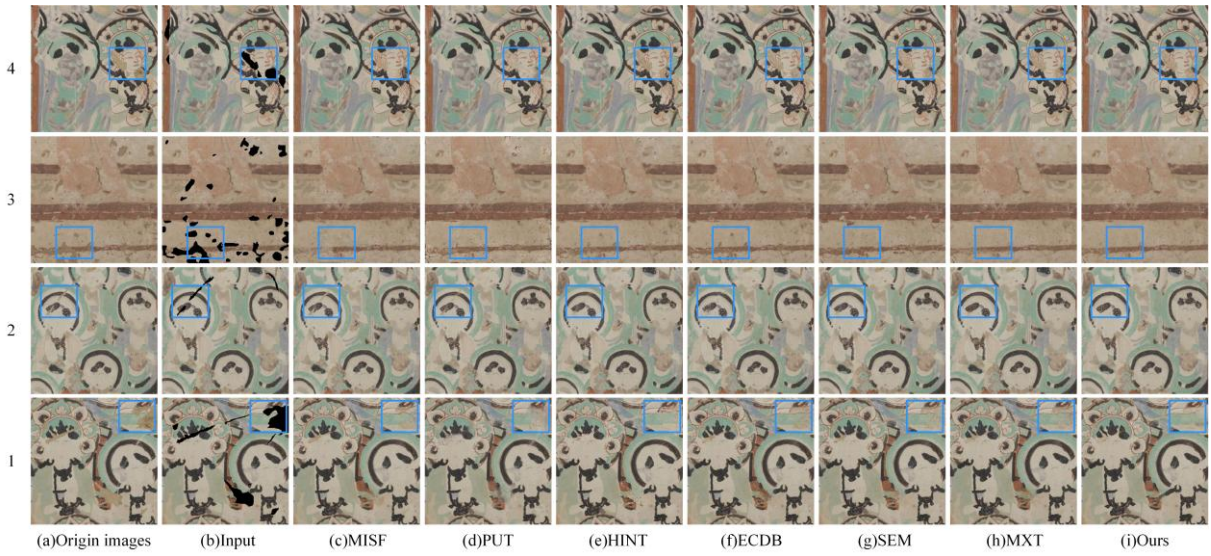


Fig.8 Comparison chart of actual damage repair experiment results

Conclusion

This paper presents a Visual Context-Aware and Spatial Receptive Field-Based (VCSR) algorithm for mural inpainting, specifically designed to address the challenges of local information loss and insufficient spatial perception in high-resolution mural restoration. The proposed method incorporates several key innovations to enhance restoration performance, including the Residual State Space Module (RSSB), a prompt vector-based local detail compensation strategy, a multi-scale deformable sliding window Transformer (MSDA), and a Spatially Perceptive Feedforward Network (SPFN). These modules collectively improve the algorithm's ability to capture spatial dependencies, model long-range relationships, and interact effectively

with local regions, leading to significantly enhanced restoration realism and visual quality.

Experimental results demonstrate the superiority of VCSR over existing methods in terms of both objective metrics and subjective visual assessments. In particular, ablation studies on the Dunhuang mural dataset show that VCSR outperforms the baseline model across multiple evaluation criteria, including improvements in Peak Signal-to-Noise Ratio (PSNR) ranging from 0.053 to 0.269, Structural Similarity Index (SSIM) gains of 0.006 to 0.007, reductions in L1 error between 0.001 and 0.003, and LPIPS reductions from 0.004 to 0.009. These results underscore the effectiveness of the VCSR algorithm in enhancing detail recovery, texture consistency, and structural preservation during mural restoration.

Despite these advances, certain limitations remain. In particular, the algorithm does not fully account for the artistic styles and historical-cultural contexts of mural creation. Since murals from different historical periods exhibit distinctive characteristics in color, texture, structure, and style, future research should explore how to integrate these cultural and artistic elements into the restoration process. Such an approach would not only improve the cultural fidelity of the results but also enhance their artistic authenticity, contributing to more holistic and meaningful heritage conservation.

Data Availability
The datasets used or analysed during the current study available from the corresponding author on reasonable request.

Code Availability
No datasets were generated or analysed during the current study.

Acknowledgements
The research was supported by the Gansu Provincial Joint Research Fund (No. 25JRRA1152); the Central Government-guided Local Science and Technology Development Fund Project (No. 25ZYJN001); and the Natural Science Foundation of Gansu Province (No. 23JRRA796).

Author contributions
All authors reviewed the manuscript. LY: Designed the study, wrote the main manuscript, and implemented the coding. LZM: Conducted the experiments and revised the manuscript. HWJ: Provided guidance, supervised the research, and reviewed the manuscript.

Funding
The paper was prepared without the assistance of any grants, funding, or other sources, as declared by the authors.

Competing interests
The authors declare that they have no competing interests.

References

[1] Lan LR, Sang LJ. Digital protection and practice of ancient murals. Art Educ 2020;5:170–3.

[2] Nazeri K, Ng E, Joseph T, et al. EdgeConnect: structure guided image inpainting using edge prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops; 2019. p. 0–0.

[3] Liu H, Wan Z, Huang W, et al. PD-GAN: probabilistic diverse GAN for image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 9371–81.

[4] Zheng C, Song G, Cham TJ, et al. High-quality pluralistic image completion via code shared VQGAN. arXiv preprint arXiv:2204.01931, 2022.

[5] Ding X, Zhang X, Han J, et al. Scaling up your kernels to 31x31: revisiting large kernel design in CNNs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 11963–75.

[6] Dai T, Cai J, Zhang Y, et al. Second-order attention network for single image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019. p. 11065–74.

[7] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 7132–41.

[8] Song Y, Yang C, Lin Z, et al. Contextual-based image inpainting: infer, match, and translate. In:

Proceedings of the European Conference on Computer Vision (ECCV); 2018. p. 3–19.

[9] Liu H, Jiang B, Xiao Y, et al. Coherent semantic attention for image inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019. p. 4170–9.

[10] Li W, Lin Z, Zhou K, et al. MAT: mask-aware transformer for large hole image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 10758–68.

[11] Wan Z, Zhang J, Chen D, et al. High-fidelity pluralistic image completion with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 4692–701.

[12] Dong Q, Cao C, Fu Y. Incremental transformer structure enhanced image inpainting with masking positional encoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 11358–68.

[13] Liang J, Cao J, Sun G, et al. SwinIR: image restoration using Swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 1833–44.

[14] Chen X, Wang X, Zhou J, et al. Activating more pixels in image super-resolution transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. p. 22367–77.

[15] Liu P, Guo H, Dai T, et al. CALF: aligning LLMs for time series forecasting via cross-modal fine-tuning. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2025. 39(18):18915–23.

[16] Gu A, Goel K, Ré C. Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396, 2021.

[17] Gu A, Dao T. MAMBA: linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752, 2023.

[18] Hu VT, Baumann SA, Gui M, et al. ZIGMA: a DIT-style zigzag MAMBA diffusion model. In: European Conference on Computer Vision. Cham: Springer Nature Switzerland; 2024. p. 148–66.

[19] Islam MM, Hasan M, Athrey KS, et al. Efficient movie scene detection using state-space transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. p. 18749–58.

[20] Qin S, Wang J, Zhou Y, et al. MAMBAVC: learned visual compression with selective state spaces. arXiv preprint arXiv:2405.15413, 2024.

[21] Xing Z, Ye T, Yang Y, et al. SEGmamba: long-range sequential modeling MAMBA for 3D medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland; 2024. p. 578–88.

[22] Chen S, Atapour-Abarghouei A, Zhang H, et al. MXT: MAMBA X transformer for image inpainting. arXiv preprint arXiv:2407.16126, 2024.

[23] Li X, Guo Q, Lin D, et al. MISF: multi-level interactive siamese filtering for high-fidelity image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 1869–78.

[24] Tian H, Zhang L, Li S, et al. Pyramid-VAE-GAN: transferring hierarchical latent variables for image inpainting. Comput Vis Media 2023;9(4):827–41.

[25] Chen S, Atapour-Abarghouei A, Shum HPH. HINT: high-quality inpainting transformer with mask-aware encoding and enhanced attention. IEEE Trans Multimed 2024;26:7649–60.

[26] Yue C, Peng Z, Ma J, et al. Enhanced control for diffusion bridge in image restoration. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE; 2025. p. 1–5.

[27] Chen S, Zhang H, Atapour-Abarghouei A, et al. SEM-Net: efficient pixel modeling for image inpainting with spatially enhanced SSM. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE; 2025. p. 461–71.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.docx
- Table2.docx