IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 3, June, 2026, pp. 1257-1270
Journal homepage: http://iieta.org/journals/ts

Generative Artificial Intelligence-Empowered Image Reconstruction and Visual Information Completion

Cheng Wang

General Education Department, Hubei Institute of Fine Arts, Wuhan 430205, China

Corresponding Author Email: wangcheng@hifa.edu.cn

https://doi.org/10.18280/ts.430314

Received: 10 December 2025
Revised: 15 May 2026
Accepted: 19 May 2026
Available online: 30 June 2026

Keywords:
image reconstruction, visual information completion, latent diffusion model, sparse self-attention, geometric prior guidance

ABSTRACT

The iterative advancement of generative artificial intelligence has established diffusion models as the predominant technical paradigm for visual corruption repair. Nevertheless, existing generative reconstruction methods remain constrained by notable deficiencies, impeding the realization of high-fidelity and physically plausible image completion outcomes. To address the aforementioned technical bottlenecks, a structure-texture dual-branch collaborative generation network was proposed, enabling high-precision and robust image reconstruction and visual information completion. The proposed method adopts a latent diffusion model as the foundational backbone and incorporates a context-adaptive discrepancy dynamic sampling strategy, achieving dynamic adaptation and seamless blending between generated regions and original image content. A lightweight geometric estimation module was designed, which leverages a self-supervised learning paradigm to learn depth and normal geometric features from unannotated images; cross-modal embedding of three-dimensional prior information was accomplished via epipolar geometric constraints, effectively rectifying physical deviations in perspective and occlusion within reconstructed content. Concurrently, the Transformer attention mechanism was optimized through the construction of a hybrid sparse self-attention and energy-gated feature alignment module, reducing the computational complexity of attention to near-linear levels while preserving representational capacity and suppressing feature-weight noise induced by image masks. Furthermore, a multi-scale progressive reconstruction mechanism was introduced, employing a hierarchical iterative strategy that encompasses coarse-grained global semantic modeling and fine-grained local texture refinement, thereby holistically balancing the integrity of global scene layout with the authenticity of detailed textures. Comprehensive comparative experiments and ablation validations were conducted on publicly available benchmark datasets. Results demonstrated that the proposed method achieved superior performance over existing state-of-the-art algorithms across core evaluation metrics. Consistently excellent stability and generalization capability were exhibited under extreme image degradation scenarios and cross-domain transfer tasks. Through multi-module collaborative innovations, a comprehensive modeling framework for generative image completion was advanced, effectively overcoming the inherent limitations of conventional approaches and providing a viable technical reference for the practical deployment of high-precision visual reconstruction technologies.

1. INTRODUCTION

Image reconstruction and visual information completion constitute foundational and core tasks in the field of computer vision, aiming to remedy visual defects such as occlusions, pixel loss, and transmission distortions. By leveraging the semantic and textural distribution patterns of complete regions, the underlying scene information is restored, rendering these tasks essential prerequisites for a wide spectrum of high-level intelligent vision applications [1, 2]. With the rapid development of digital imaging technologies and the artificial intelligence industry, high-precision image completion has emerged as an indispensable enabling technology across multiple domains. In remote sensing for Earth observation, cloud occlusion and sensor-induced missing data can be effectively repaired, thereby ensuring data integrity for land surveying, environmental monitoring, and disaster analysis [3, 4]. In medical image diagnosis, imaging noise, local artifacts, and lesion-region occlusions can be eliminated, enhancing the accuracy and reliability of medical visual analysis [5, 6]. In digital media and intelligent content creation, this technology provides core support for applications such as super-resolution, scene extrapolation, and legacy photograph restoration [7, 8]. Moreover, image completion serves as a critical means for repairing perceptual blind spots and refining scene information in machine vision tasks, including autonomous driving and scene perception [9, 10]. Conventional image completion methods rely on handcrafted priors such as texture matching, sparse coding, and exemplar-based copying, which are capable of handling only simple textural scenarios and fail to

comprehend deep semantic structures. Consequently, their performance in complex scenes with large-area missing regions is markedly limited [11, 12]. The revolutionary advancement of generative artificial intelligence has fundamentally transformed the technical paradigm of visual information completion. In comparison with traditional generative adversarial networks and Vision Transformer-based approaches, diffusion models have become the predominant research direction for image reconstruction tasks, owing to their superior semantic fitting capacity, enhanced generation stability, and more refined detail modeling capabilities [13, 14]. Nevertheless, existing generative image completion techniques remain constrained by multidimensional technical deficiencies, rendering it difficult to simultaneously satisfy contextual consistency, physical and geometric plausibility, computational efficiency, and global-local adaptability [15, 16]. In light of these challenges, the key technical bottlenecks in generative image reconstruction are addressed in this study, and a high-precision, high-robustness visual completion generation framework is constructed. The associated research not only enriches the theoretical modeling system of generative low-level vision tasks and compensates for inherent deficiencies in current techniques, but also offers practically viable solutions for industrial-grade high-fidelity image restoration and cross-scenario visual information completion, thereby possessing significant theoretical research value and substantial engineering application potential.

Through years of iterative technological development, deep learning methodologies for image reconstruction and visual completion have converged into three predominant technical paradigms: generative adversarial networks, Vision Transformers, and diffusion models. Across these paradigms, continuous improvements to completion performance have been achieved through advancements in semantic modeling, feature extraction, and global correlation modeling, significantly enhancing the restoration quality of degraded images [17, 18]. Nevertheless, each of the three methodological families is encumbered by inherent technical limitations, rendering them incapable of simultaneously satisfying the demands for high precision, high robustness, and high physical consistency in complex scenes. Four critical and pressing challenges persist in existing generative image completion techniques. The first challenge concerns the contextual consistency rupture during the diffusion generation process. Prevailing diffusion-based completion methods adopt a fragmented generation strategy in which known regions are fixedly preserved while unknown regions are independently generated. Pixel-level constraints on valid regions are imposed only at the final output stage, without enforcing contextual distribution alignment supervision throughout the full reverse diffusion iterative process. Consequently, the textures, colors, and structural features generated within missing regions fail to achieve natural integration with the intact regions of the original image, ultimately manifesting as visual distortions, including abrupt boundaries, stylistic fragmentation, and semantic mismatches [19, 20].

The second challenge pertains to the absence of three-dimensional geometric structure priors. Existing generative models are trained and inferred solely on the basis of two-dimensional Red, Green, and Blue (RGB) pixel features, relying entirely on apparent image information for reconstruction without incorporating geometric priors such as depth and normal vectors. The intrinsic perspective laws and object occlusion relationships of real-world scenes cannot be learned by the model, resulting in reconstructed images that appear superficially plausible in two-dimensional visual perception yet ubiquitously suffer from physical inconsistencies, including spatial structural distortion, perspective scaling imbalances, and erroneous occlusion ordering. These deficiencies substantially compromise the authenticity and credibility of image reconstruction [21, 22]. The third challenge involves feature noise and computational performance bottlenecks inherent in attention mechanisms. Global attention in Vision Transformers serves as the core mechanism for modeling long-range semantic correlations. However, the localized information deprivation induced by image masks severely disrupts the allocation logic of attention weights, giving rise to weight distortion and spurious feature noise that critically impair feature alignment precision. Simultaneously, the inherently high computational complexity of standard global attention imposes severe constraints on the deployment of models in high-resolution image reconstruction scenarios, manifesting as prohibitive computational overhead, slow inference speeds, and elevated hardware costs [23, 24]. The fourth challenge is the modeling conflict between global semantics and local details. Reconstruction paradigms with fixed resolution are ill-suited to the concurrent modeling demands of multi-scale visual information. Low-resolution reconstruction approaches effectively preserve global scene layout and semantic integrity, yet inevitably sacrifice high-frequency textural details, producing overly smooth reconstructed outputs. Conversely, high-resolution reconstruction faithfully recovers fine textural features, but is prone to local artifacts and structural distortions, while incurring a steep escalation in computational cost. The holistic integrity of global semantics and the verisimilitude of local details cannot be jointly optimized under such fixed-resolution schemes [25].

To address the aforementioned four core technical challenges in image reconstruction and visual completion, a structure-texture dual-branch collaborative generation network is systematically constructed in this study, through which the full-process modeling capability of generative completion is optimized via multi-module collaborative innovations. The core research contributions are fourfold. A context-adaptive discrepancy dynamic sampling strategy is proposed, by which the conventional outcome-oriented consistency constraint is transformed into a dynamic supervision mechanism spanning the entire diffusion iterative process, thereby ensuring natural contextual integration of image content from the perspective of the generative procedure. A lightweight geometric estimation module is designed, through which depth and normal features are estimated online via a self-supervised learning paradigm, and three-dimensional geometric priors are incorporated at no additional cost through the integration of epipolar geometric constraints, effectively resolving the issue of physical structural inconsistency in reconstructed content. A hybrid sparse self-attention and energy-gated feature alignment module is constructed, through which the computational complexity of attention is substantially reduced via decoupling of spatial and channel attention along with sparse selection mechanisms, while feature noise induced by masks is effectively suppressed. A multi-scale progressive reconstruction mechanism is introduced, through which the modeling conflict between global layout and local details is reconciled via a hierarchical generation paradigm encompassing coarse-grained global semantic construction

and fine-grained local texture refinement, thereby comprehensively enhancing the quality and robustness of image completion in complex scenes.

The remainder of this study is organized as follows. In Chapter 2, the overall framework of the proposed structure-texture dual-branch collaborative generation network, the encoding and decoding logic of the dual-branch features, the design principles of each core innovation module, and the loss functions and training strategies are elaborated in detail. In Chapter 3, multiple sets of comparative experiments, ablation studies, and robustness tests are designed, through which the superiority of the proposed method and the effectiveness of each module are comprehensively validated via quantitative metrics, qualitative visualization, and efficiency analysis. In Chapter 4, the research findings of the entire study are summarized, the core conclusions are distilled, and future research directions and application scenarios are prospected in light of the limitations of existing studies.

2. STRUCTURE-TEXTURE DUAL-BRANCH COLLABORATIVE GENERATION NETWORK

2.1 Overall framework and feature encoding-decoding pipeline

The proposed structure-texture dual-branch collaborative generation network is built upon a latent diffusion model as its foundational architecture, through which the reconstruction and refinement of missing image information are accomplished in a low-dimensional latent space. By means of decoupled modeling of structural and textural features along with a collaborative fusion mechanism, the coupling between structural distortion and textural artifacts inherent in conventional generative methods is effectively resolved. An overview of the overall network architecture and the feature encoding-decoding pipeline of the structure-texture dual-branch collaborative generation network is presented in Figure 1. The inputs to the network are the original degraded image and a binary mask matrix, which are respectively defined as the image matrix $I \in \mathrm{R}^{H\times W\times 3}$ and the mask matrix $M \in \{0,1\}^{H\times W}$, where H and W denote the height and width of the image, respectively. To reduce the modeling overhead associated with high-resolution images, a variational autoencoder is employed for feature compression, with a downsampling factor set to 8. The encoder maps the pixel-space image to a dimensionally compressed latent feature $z = \mathrm{E}(I) \in \mathrm{R}^{h\times w\times c}$, where $h = H/8$ and $w = W/8$ represent the spatial dimensions of the latent feature, and the channel dimension is $c = 4$. Simultaneously, the original mask matrix is downsampled in a corresponding manner, yielding a mask feature M_{lat} adapted to the latent space, thereby achieving precise alignment of missing-region information at the feature level.

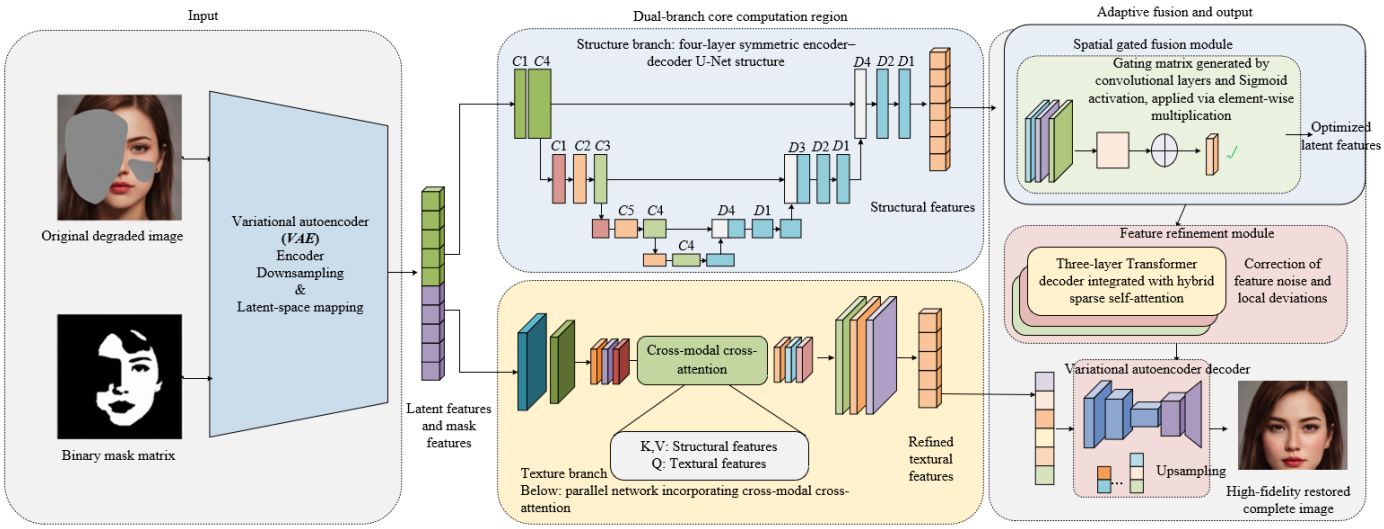

Figure 1. Overall network architecture and feature encoding-decoding pipeline of the structure-texture dual-branch collaborative generation network

A dual-branch parallel encoding-decoding architecture is constructed within the network, through which global structural modeling and local texture generation are respectively accomplished, thereby enabling differentiated representation of the two categories of visual features. The structure branch adopts a four-layer symmetric encoder-decoder U-Net architecture, taking as input the combined features obtained by channel-wise concatenation of latent features and mask features, and is dedicated to extracting image edge contours, spatial layout, and geometric structure information, outputting structural features F_s with dimensions of $h\times w\times 64$. During training, a Sobel gradient $L1$ loss is imposed on this branch to constrain its learning, thereby reinforcing the model's capacity for capturing image structural features and ensuring the overall spatial plausibility of the reconstructed content. The texture branch incorporates cross-modal cross-attention units at the decoder stage, where query matrices are constructed from the branch's own textural features, while key and value matrices are constructed by reusing the structural features output from the structure branch. Through cross-layer feature interaction, global guidance of texture generation by structural priors is achieved, ultimately yielding refined textural features F_t of the same dimensionality, ensuring that the texture generation process strictly conforms to the overall structural logic of the image.

To achieve adaptive fusion of structural and textural features, a spatial gated fusion module is introduced into the model, through which an adaptive weight distribution is learned via convolutional layers and a Sigmoid activation function, yielding a spatial gating matrix G. This computation

is formulated as:

$$G = \sigma(Conv([F_s;F_t])) \tag{1}$$

where, σ denotes the Sigmoid activation function, *Conv* denotes the convolution operation, and [;] denotes the channel-wise concatenation operation. Adaptive fusion of the dual features is accomplished via the gating matrix, yielding the optimized latent feature:

$$\hat{z} = G \odot F_s + (1 - G) \odot F_t \tag{2}$$

where, $\odot$ denotes the element-wise multiplication operation. The fused latent feature is then fed into a three-layer Transformer decoder integrated with hybrid sparse self-attention for global feature refinement, through which feature noise and local deviations are corrected. Finally, feature upsampling and pixel-space reconstruction are performed via the variational autoencoder decoder, yielding the restored complete image $\hat{I}=D(\hat{z})$, through which bi-directional optimality of global structural regularity and local textural authenticity is achieved.

2.2 Context-adaptive dynamic diffusion sampling mechanism

An overview of the core innovative components and the multi-modal alignment micro-structure is presented in Figure 2. The latent diffusion model accomplishes the restoration of missing latent features through multiple rounds of reverse Gaussian denoising. The standard reverse generation process conforms to a Gaussian conditional distribution, mathematically formulated as:

$$p_\theta(z_{t-1}|z_t) = N(z_{t-1};\mu_\theta(z_t,t),\Sigma_\theta(z_t,t)) \tag{3}$$

where, t denotes the current time step of the reverse diffusion iteration; z_t denotes the latent feature map at step t; μ_θ and Σ_θ respectively denote the distribution mean and variance predicted by the network; and N denotes the multivariate Gaussian distribution. Conventional sampling strategies employ a fixed and uniform time step across the entire iterative cycle, rendering them incapable of adaptively distinguishing between smooth image regions and structural boundaries where abrupt transitions occur. The distributional deviation between known pixels and pixels to be generated at the missing-region boundaries cannot be progressively corrected throughout the iterative process, ultimately resulting in pronounced visual discontinuities between the restored region and the original image content. In this study, a context-adaptive discrepancy metric is designed, through which local distributional deviations are computed at each spatial location of the feature map during every denoising iteration, thereby guiding the dynamic adjustment of sampling step sizes and iteration counts, and embedding contextual consistency constraints throughout the complete diffusion generation pipeline.

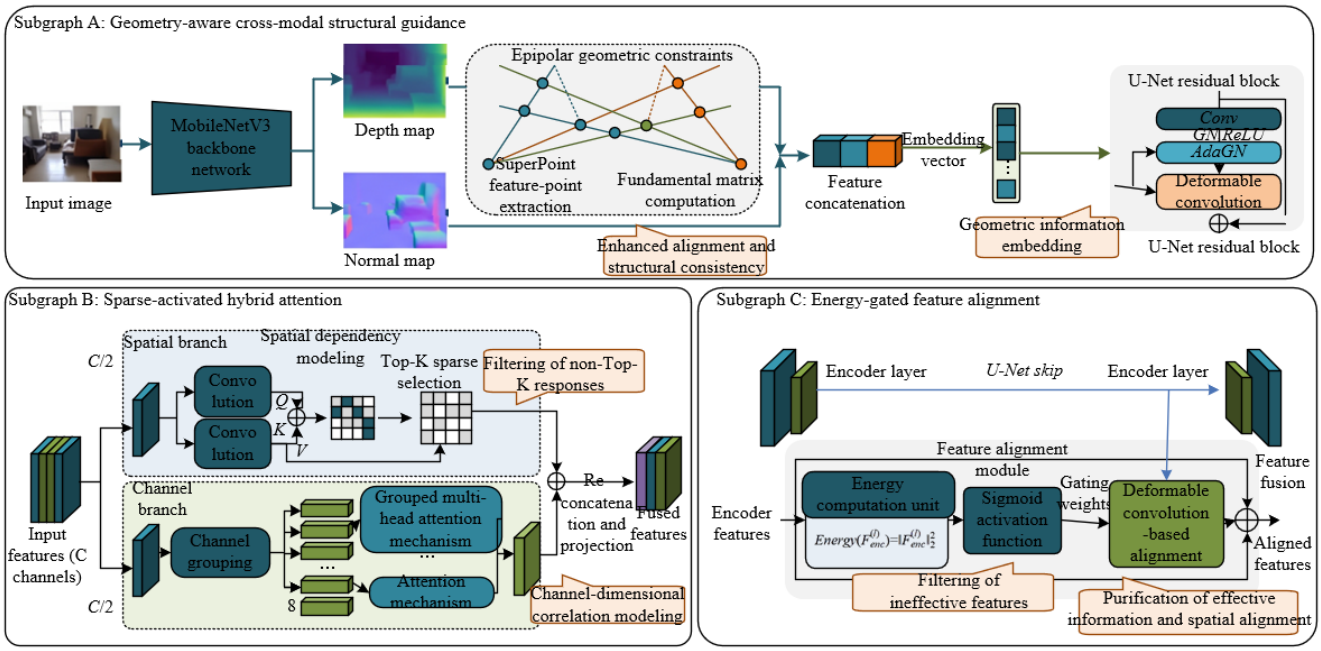

Figure 2. Core innovative components and multi-modal alignment micro-structure of the proposed framework

For any spatial coordinate p on the latent feature map, a 3×3 local neighborhood $N(p)$ centered at that coordinate is selected, with all pixels within the neighborhood collectively denoted as q. The Gaussian distribution parameters corresponding to intact pixels within the neighborhood are denoted as μ_{known} and Σ_{known}, while the distribution parameters corresponding to pixels to be generated in the missing region are denoted as $\mu_{unknown}$ and $\Sigma_{unknown}$. The mechanism simultaneously incorporates the Kullback-Leibler divergence and the mean squared deviation to quantify the degree of mismatch between the two classes of feature distributions. A weighted summation is performed to obtain the local discrepancy loss at a single pixel position at step t, with the complete formulation given as:

$$L_{CAD}^{(t)}(p) = \sum_{q \in N(p)} [\, KL(N(\mu_{known}(q),\Sigma_{known}(q)) \| N(\mu_{unknown}(q),\Sigma_{unknown}(q))) + \lambda \| \mu_{known}(q) - \mu_{unknown}(q) \|_2^2] \tag{4}$$

where, *KL* denotes the Kullback-Leibler divergence operator, used to quantify the overall distributional shape difference between the two Gaussian distributions; λ denotes a hyperparameter for balancing the weights of the distributional shape constraint and the mean-distance constraint; and $\| \|_2^2$ denotes the squared L_2-norm operation. This loss function is computed synchronously with the diffusion time steps, enabling step-by-step and pixel-wise consistency supervision throughout the generation process, thereby avoiding the delayed correction issue associated with constraints imposed only at the final output stage.

Based on the local discrepancy loss, a spatially adaptive adjustment factor is constructed to determine whether refinement sampling iterations are required at a given position, formulated as:

$$\alpha^{(t)}(p) = \sigma(\gamma \cdot L_{CAD}^{(t)}(p) - \beta) \tag{5}$$

where, σ denotes the Sigmoid activation function, which maps the computed result to the interval [0,1]; γ denotes a loss scaling coefficient, which amplifies the difference magnitude across different regions; and β denotes a global bias threshold, which controls the trigger sensitivity of the refinement operation. A decision threshold of 0.5 is set in this study. When the adjustment factor exceeds this threshold, a significant feature distribution shift is indicated in the local region, necessitating additional refinement denoising. Conversely, when the value falls below the threshold, the region is determined to be texturally smooth, and the existing sampling precision is deemed sufficient for reconstruction requirements.

When a local region is determined to require refinement correction, a 5 × 5 effective range is demarcated centered at coordinate p , and refinement iterations are performed exclusively within this range. The refinement process employs a scaled time step $\eta \cdot \Delta t$, with the scaling coefficient fixed at 0.3, while two additional independent denoising iterations are executed, with the total number of refinement iterations denoted as $N_{refine} = 2$. The smaller iteration step size enables fine-grained adjustment of the latent feature distribution, and the two repeated iterations sufficiently smooth the distribution discontinuities between known and generated regions. For pixels outside the effective range, the original sampling step size is retained without additional computational overhead. In this manner, computational cost is incurred exclusively for local regions with complex structures and significant distribution deviations, while uniformly textural regions bypass redundant iterations, thereby controlling overall inference computational load while ensuring boundary naturalness.

2.3 Geometry-aware cross-modal structural guidance

The geometry-aware cross-modal structural guidance module is dedicated to injecting three-dimensional geometric priors into the image reconstruction process, thereby resolving the issues of perspective disorganization and occlusion logic contradictions inherent in two-dimensional pixel-driven generation. The core of this module lies in the deep synergy between a lightweight geometric estimation module and a cross-modal feature fusion mechanism. The module is composed of three components: geometric information generation, feature embedding, and multi-view constraints, through which precise geometric guidance is achieved without reliance on ground-truth three-dimensional annotations. The lightweight geometric estimation module adopts MobileNetV3 as its backbone network, complemented by a three-layer deconvolutional structure, forming an encoder-decoder architecture. Given the input image, the depth map and surface normal map are simultaneously estimated online, respectively denoted as $D \in \mathrm{R}^{H \times W}$ and $N \in \mathrm{R}^{H \times W \times 3}$, where H and W denote the height and width of the input image. The depth map records the relative spatial distance of each pixel, while the normal map describes the direction of the surface normal vector at each pixel.

The module is trained under a self-supervised learning paradigm, where a photometric loss function is constructed via cross-frame reprojection error, enabling the acquisition of geometric estimation capability without annotations. During training, adjacent frames in a video sequence are randomly selected as the source frame and the target frame. The source frame is projected into the target frame coordinate system based on the currently predicted depth and camera intrinsics, and the pixel-wise photometric error between the projected image and the target frame is computed. Through this error, the parameters of the geometric estimation module are optimized in a backward manner, circumventing reliance on large-scale annotated three-dimensional datasets. After geometric information generation, the concatenated features of the depth map and normal map are projected into a geometrically embedded vector $e_{geo} \in \mathrm{R}^{h \times w \times 128}$ via a two-layer multi-layer perceptron, where $h = H/8$ and $w = W/8$ are consistent with the spatial dimensions of the latent feature map, and 128 denotes the channel dimension of the embedding vector, ensuring seamless integration of geometric information with subsequent generative features.

The geometric embedding is integrated into the generation network through a deep fusion approach, enabling geometric priors to guide the entire diffusion process. In each residual block of the U-Net, deformable convolution is first applied to the geometric embedding e_{geo} for feature alignment, where pixel-level offsets are adaptively adjusted through learning to rectify spatial misalignment between geometric features and latent generative features, ensuring that geometric information precisely matches the structural distribution of the current generated region. The aligned geometric embedding and the current time-step embedding e_t are added channel-wise and jointly fed into an adaptive group normalization layer, through which the intermediate features of the residual block are modulated via learned scaling factors and biases. This computational process is formulated as:

$$F_{mod} = \gamma \cdot (F_{res} - \mu)/\sigma + \beta \tag{6}$$

where, F_{res} denotes the input feature of the residual block; μ and σ denote the mean and standard deviation of the features within each group; and γ and β denote the adaptively learned parameters, which are generated by a fully connected layer from the fused features of the geometric embedding and the time-step embedding. In this manner, the geometric constraints are enabled to dynamically adapt to the feature distributions at different generation stages, achieving deep binding between geometric priors and the generative process.

For multi-view input scenarios, epipolar geometric constraints are further introduced into the module to reinforce the physical consistency of the three-dimensional structure, through which a composite geometric loss encompassing epipolar error and disparity consistency is constructed. First,

the SuperPoint algorithm is employed to extract keypoints from the two-view images and perform matching, and the fundamental matrix $F \in \mathrm{R}^{3\times3}$ between the two views is solved via the eight-point algorithm. This matrix encodes the projective geometric relationship between the two views and satisfies the epipolar constraint $x_2^{\mathrm{T}} F x_1 = 0$, where x_1 and x_2 denote the homogeneous coordinates of corresponding points in the two views. Based on the fundamental matrix and the estimated depth, the geometric loss function is defined as:

$$L_{geo} = \sum_i \| x_2^{(i)\mathrm{T}} F x_1^{(i)} \|_1 + \| d_{warp} - d_{epi} \|_1 \tag{7}$$

where, the first term denotes the epipolar constraint loss, which iterates over all matched point pairs $(x_1^{(i)}, x_2^{(i)})$ to compute the epipolar error, ensuring that corresponding points satisfy the geometric projection relationship; the second term denotes the disparity consistency loss, where d_{warp} denotes the disparity map obtained by transforming the estimated depth map through camera parameters, and d_{epi} denotes the disparity map derived from epipolar rectification. The reliability of depth estimation is guaranteed by enforcing consistency between these two disparity representations. This composite loss is jointly optimized with the reconstruction loss of the generation network, ensuring that the completed content is not only visually coherent in two-dimensional appearance but also physically consistent with the perspective and occlusion regularities of the real world in the three-dimensional geometric domain.

2.4 Sparse-activated hybrid attention and energy-gated feature alignment

To balance the capacity for long-range visual feature correlation modeling against computational resource overhead, while simultaneously alleviating the issue of attention weight distortion induced by image corruption masks, a sparse-activated hybrid attention mechanism is constructed in this study, through which decoupled modeling of spatial-dimension and channel-dimension correlations is achieved. This mechanism performs differentiated attention computation on input features, discarding the redundant computation paradigm of standard global dense attention in Transformers. For any given input latent feature X, the network first generates query, key, and value matrices via linear transformations, sequentially denoted as Q, K, and V. The total feature channels are equally divided into two independent branches, with the channel dimension of each branch taking half of the original dimensionality, respectively designated for spatial correlation modeling and channel correlation modeling, thereby enabling separated representation of the two types of feature dependencies and avoiding mutual interference between the two dimensions. An overview of the multi-scale progressive diffusion generation and dynamic sampling pipeline is presented in Figure 3.

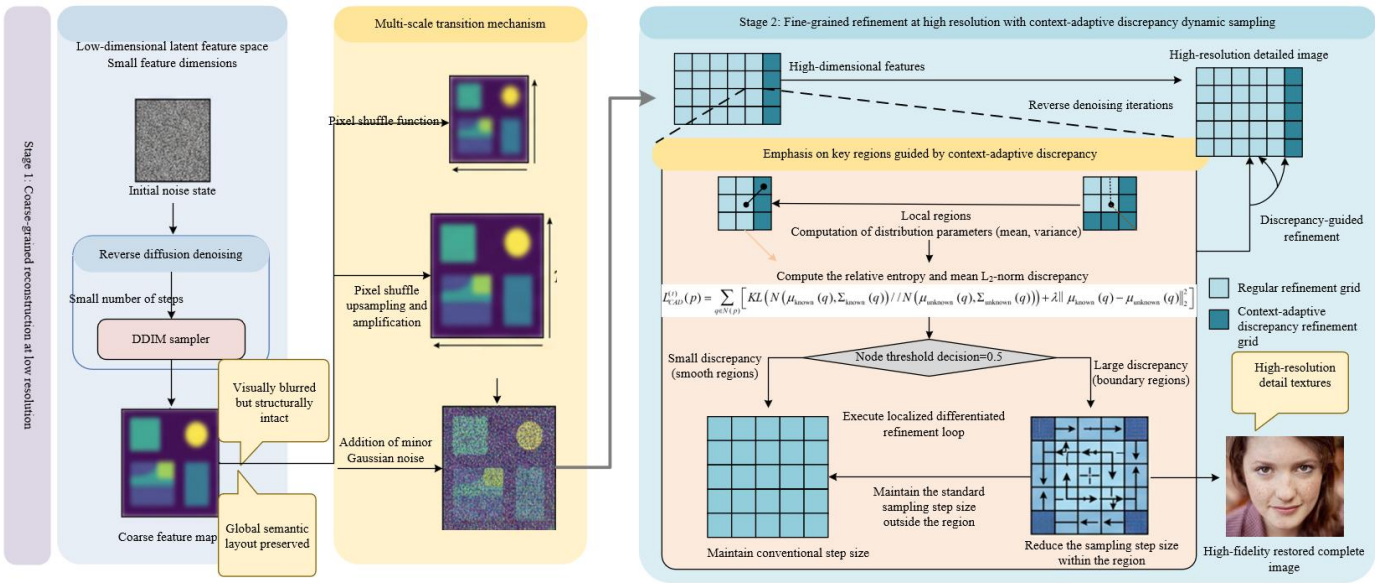

Figure 3. Multi-scale progressive diffusion generation and dynamic sampling pipeline

In the spatial branch, a sparse attention selection strategy is adopted, through which effective correlative features in the image pixel space are focused upon, with the attention response matrix computed as:

$$A = \frac{Q_s K_s^{\mathrm{T}}}{\sqrt{d/2}} \tag{8}$$

where, Q_s and K_s denote the query and key features of the spatial branch, respectively, and $d/2$ denotes the channel dimension of each branch, through which numerical normalization is achieved via the square root of the dimensionality, ensuring gradient propagation stability. A Top-K selection operation is performed on the response values of each row of the attention matrix, with the number of retained nodes set to $K = \lfloor N/4 \rfloor$, where N denotes the total number of pixels in the feature space. After selection, only the top-ranked attention weights with the highest similarity are retained, while all remaining weights are uniformly set to negative infinity, followed by weight normalization via the Softmax function, yielding the refined spatial correlation features $Z_{spatial}$. In the channel branch, a grouped multi-head attention mode is adopted, where the query, key, and value features along the channel dimension are equally divided into 8 independent subgroups. Attention computation is performed independently within each subgroup, followed by feature

concatenation, yielding channel-dimensional correlation features $Z_{channel}$. After channel-wise concatenation of the two sets of branch features, feature fusion is accomplished via linear projection, through which the computational complexity of attention is effectively reduced, optimizing the original standard attention complexity of $O(N^2 d_k + N d_k^2)$ to $O(KNd_k + Nd_k^2/G)$, thereby significantly reducing the inference overhead for high-resolution image reconstruction.

The energy-gated feature alignment module operates on the encoder-decoder skip connections of the network, through which adaptive screening and precise alignment of cross-level features are achieved, and representational noise introduced by ineffective features in missing regions is suppressed. In this module, the energy distribution of multi-level encoder features is computed to quantify the amount of effective information in the features, with the feature energy formulated as:

$$Energy(F_{enc}^{(l)}) = \|F_{enc}^{(l)}\|_2^2 \tag{9}$$

where, $F_{enc}^{(l)}$ denotes the output feature of the encoder at the l-th layer, and the squared L_2-norm operation effectively characterizes the feature response intensity. The normalized feature energy is then fed into the Sigmoid activation function to generate adaptive feature gating weights:

$$Gate = Sigmoid(E\text{-}\tau) \tag{10}$$

where, E denotes the normalized feature energy matrix, and τ denotes the energy threshold parameter used to distinguish effective features from noise features. After the gating weights are applied to screen the encoder features, spatial position-adaptive alignment is accomplished via deformable convolution, through which spatial offset errors between encoder and decoder features are eliminated. Finally, cross-level feature fusion is performed, with the fusion rule satisfying:

$$F_{align}^{(l)} = F_{dec}^{(l)} + Gate \cdot Proj(F_{enc}^{(l)}) \tag{11}$$

where, $F^{(l)}{}_{dec}$ denotes the decoder feature at the corresponding level, and $Proj$ denotes the feature projection operation. This mechanism enables adaptive filtering of ineffective features in missing regions without reliance on mask information, substantially improving the feature fusion quality of skip connections.

The hybrid attention mechanism and the energy-gated alignment module form a complementary optimization framework. The former reduces computational redundancy and corrects attention biases induced by masks at the feature modeling level, while the latter purifies cross-level effective information at the feature fusion level. Through their synergistic operation, both the efficiency bottleneck and the noise interference issues of conventional Vision Transformers in image completion tasks are effectively addressed, while the fusion precision of global semantic features and local detail features is simultaneously reinforced, providing high-quality feature representations for subsequent refined image reconstruction.

2.5 Multi-scale progressive reconstruction strategy

To alleviate the difficulty in jointly optimizing global semantic layout and local high-frequency textures under the fixed-resolution reconstruction paradigm, a multi-scale progressive reconstruction strategy is constructed in this study, through which scene structure modeling and texture detail restoration are accomplished stepwise via a coarse-to-fine hierarchical iterative generation mechanism. The strategy first performs coarse-grained reconstruction in the low-dimensional space, where the original latent feature map is compressed to dimensions of $h/2 \times w/2$, and low-frequency information restoration is accomplished via the DDIM sampler, with the total number of sampling steps set to $T_{coarse} = 50$. A linear variance scheduling scheme is adopted during the sampling process to ensure the stability and continuity of the diffusion iterative process, with priority given to fitting the overall scene structure, object contours, and global semantic distribution of the image, thereby effectively circumventing the structural distortion and layout disorganization issues prone to occurrence in high-dimensional iterative processes, and yielding a low-resolution latent feature $\hat{z}_{low}$ with complete scene logic.

After global structural modeling is accomplished, texture detail optimization is achieved through progressive scale transformation and fine-grained iteration. A pixel shuffle algorithm combined with a 3×3 convolutional layer is adopted for feature upsampling, through which the feature discontinuities induced by scale transformation are smoothed while the high-resolution latent space dimensions are restored, ensuring the continuity of feature-space dimensionality switching. To activate the detail generation capability of high-dimensional features and enhance texture diversity, a small amount of Gaussian noise is added to the upsampled features, with the noise distribution satisfying:

$$\epsilon \sim N(0, 0.2^2 I) \tag{12}$$

where, N denotes the Gaussian distribution, 0.2 denotes the noise standard deviation, and I denotes the identity matrix. The noise-added high-dimensional features are taken as the initial iterative features, and a refined reverse denoising process with $T_{fine} = 10$ steps is performed. During the fine-grained generation stage, the coarse-grained reconstructed features are incorporated into the generation constraints as additional cross-attention conditions, ensuring that the high-frequency texture generation process consistently adheres to the established global structure, thereby fundamentally suppressing local artifacts and textural disarray.

To further strengthen the feature constraint capability of multi-scale reconstruction, a multi-scale discriminator is introduced for comprehensive adversarial calibration of the hierarchical generation results. The multi-scale discriminator is composed of three structurally identical PatchGAN sub-networks, which perform feature discrimination on the original-resolution image, the 2× downsampled image, and the 4× downsampled image, respectively, covering multi-dimensional feature hierarchies including global scene, meso-scale structure, and micro-scale texture. An equal-weight fusion strategy is adopted for the adversarial losses across the three scales, yielding the final multi-scale adversarial loss function:

$$L_{adv} = \frac{1}{3}\left(L_{adv}^0 + L_{adv}^2 + L_{adv}^4\right) \tag{13}$$

where, L_{adv}^0, L_{adv}^2, and L_{adv}^4 denote the adversarial losses at the original scale, the 2× downsampled scale, and the 4×

downsampled scale, respectively. Through the multi-scale constraint mechanism, the global structural deviations at the coarse stage and the local textural errors at the fine stage are simultaneously calibrated, enabling synergistic optimization of global semantic integrity, structural geometric plausibility, and local textural authenticity.

2.6 Overall loss function and two-stage training strategy

A multi-objective composite loss function is adopted by the model to simultaneously constrain pixel-level reconstruction accuracy, contextual distribution consistency, three-dimensional geometric plausibility, generative distribution realism, and high-level visual perceptual features. The complete loss expression is formulated as:

$$L_{total} = \underbrace{\|I-\hat{I}\|_1 + 0.2\|\nabla I-\nabla\hat{I}\|_1}_{L_{recon}} + \lambda_1 L_{CAD} + \lambda_2 L_{geo} + \lambda_3 L_{adv} + \lambda_4 \underbrace{\sum_{j\in\{2,3,4\}} \|\phi_j(I) - \phi_j(\hat{I})\|_2}_{L_{perc}} \quad (14)$$

where, each sub-loss corresponds to a different level of generation constraint. L_{recon} denotes the base reconstruction loss, composed of a weighted combination of pixel-space $L1$ loss and gradient $L1$ loss, with the gradient term weight set to 0.2, through which edge and contour restoration effects are reinforced; λ_1 , λ_2 , λ_3 , and λ_4 denote trainable balancing coefficients for adjusting the contribution of each auxiliary loss to the overall optimization objective; L_{CAD} denotes the context-adaptive discrepancy loss, which enforces distribution matching between known and generated regions throughout the full diffusion iterative process; L_{geo} denotes the epipolar geometric loss, ensuring that the reconstructed content conforms to the physical laws of three-dimensional perspective; L_{adv} denotes the adversarial loss output by the multi-scale discriminator, optimizing the overall distributional realism of the generated images; and L_{perc} denotes the perceptual loss, where ϕ_j refers to the pooled output features of the 2nd, 3rd, and 4th layers of the VGG-19 network, through which visual perceptual similarity is enhanced by reducing the high-level feature distance between the original image and the reconstructed image.

Direct end-to-end training of the multi-module coupled network is prone to gradient conflicts and localized module convergence lag. Therefore, a two-stage progressive training pipeline is designed in this study. In the first stage, a total of 50 epochs are executed, during which the backbone parameters of the diffusion model are completely frozen, and only the structure-texture dual branches, the lightweight geometric estimation module, and the energy-gated feature alignment module are updated. Priority is given to achieving parameter convergence of the sub-modules responsible for feature decoupling, three-dimensional prior embedding, and cross-level feature fusion, thereby simplifying the difficulty of solving the multi-task synchronous optimization problem. In the second stage, 30 epochs of global fine-tuning are performed, during which all network parameters are released for end-to-end joint optimization. The AdamW optimizer is adopted for training, with the weight decay fixed at 0.01. The learning rate follows a cosine decay scheduling rule, with the initial learning rate set to 1×10^{-4}, linearly decayed to 1×10^{-6} at the end of the iteration cycle, through which smooth parameter adaptation across sub-modules is achieved, simultaneously enhancing image reconstruction accuracy and cross-scenario generalization performance.

3. EXPERIMENTS

3.1 General experimental settings

The comprehensive performance of the structure-texture dual-branch collaborative generation network in image reconstruction and visual information completion tasks was rigorously validated through a series of experiments, including quantitative evaluations, qualitative visualizations, module ablation studies, robustness tests, cross-domain generalization assessments, and computational efficiency analyses. All experiments were conducted on three publicly available standard datasets for training and evaluation: the face dataset CelebA-HQ, the general scene dataset Places2, and the street-view dataset Paris StreetView, with all images uniformly resized to a resolution of 512 × 512. During the testing phase, two categories of masks were employed. The first category comprised 12,000 irregularly shaped masks, with missing-area ratios set to 30% and 50%. The second category comprised center rectangular masks, with missing ratios set to 20% and 40%, resulting in a total of five distinct degradation testing scenarios.

Performance evaluation was conducted using four widely adopted quantitative metrics in the field: higher values of peak signal-to-noise ratio and structural similarity index indicate better reconstruction quality, while lower values of Fréchet inception distance and learned perceptual image patch similarity indicate higher perceptual realism of the generated images. Five representative algorithms were selected as comparison baselines, including the conventional generative adversarial network-based method EdgeConnect, the Transformer-based scheme Image Completion Transformer (ICT), the diffusion-based restoration models RePaint and Palette, and the recent mainstream generative inpainting network BrushNet. All competing methods were reproduced using the publicly available code provided by the authors and the default hyperparameters specified in their respective publications. Model training and inference were implemented using the PyTorch 2.0 deep learning framework, with the hardware environment equipped with four A100 80GB graphics processing units, and the training batch size uniformly set to 16.

3.2 Comprehensive quantitative and qualitative comparison with state-of-the-art methods

In this experiment, horizontal comparisons were conducted across three datasets and five mask degradation conditions. Each experimental configuration was independently run three times, and the mean and standard deviation of the four evaluation metrics were recorded. The complete quantitative results are presented in Table 1. In all tables, the optimal metric values are highlighted in bold, with the superscript ↑ indicating that higher values correspond to superior performance, and ↓ indicating that lower values correspond to superior performance.

From the quantitative data presented in Table 1, it can be observed that across all dataset-mask combinations, the proposed structure-texture dual-branch collaborative generation network consistently achieves significantly superior performance on all four evaluation metrics compared

to all baseline methods. Taking the most challenging scenario—the 50% irregular mask with the highest degradation level—as an example, compared to the suboptimal method BrushNet, the proposed method achieves a peak signal-to-noise ratio improvement of 1.65, a structural similarity index improvement of 0.052, a Fréchet inception distance reduction of 7.56, and a learned perceptual image patch similarity reduction of 0.041, with the performance gains being more pronounced under high-degradation conditions. The conventional generative adversarial network-based method EdgeConnect exhibits the poorest overall performance, being prone to structural distortions under large-area masks. The standard Transformer-based method ICT suffers from texture blurring artifacts. The existing diffusion-based inpainting methods RePaint and Palette maintain reasonable performance only under mild degradation scenarios, but exhibit noticeable contextual discontinuities when confronted with complex large-area occlusions. BrushNet, as the most competitive baseline in recent years, incorporates structural constraints but lacks three-dimensional geometric priors and dynamic sampling mechanisms, resulting in insufficient perspective consistency in the reconstructed images. Figure 4 shows the structure-texture dual-branch collaborative generation process for image reconstruction and visual information completion.

Table 1. Average quantitative metrics (mean ± standard deviation) of different algorithms across multiple datasets and mask conditions

Dataset	Mask Type	Metric	EdgeConnect	Image Completion Transformer (ICT)	RePaint	Palette	BrushNet	Structure-Texture Dual-Branch Collaborative Generation Network (Proposed)
CelebA-HQ	Rectangular 20%	Peak signal-to-noise ratio↑	24.13 ± 0.32	25.67 ± 0.28	26.81 ± 0.24	27.05 ± 0.21	27.42 ± 0.19	28.76 ± 0.16
		Structural similarity index↑	0.821 ± 0.013	0.854 ± 0.011	0.876 ± 0.009	0.882 ± 0.008	0.891 ± 0.007	0.924 ± 0.005
		Fréchet inception distance↓	21.47 ± 1.26	16.82 ± 1.04	13.54 ± 0.87	12.18 ± 0.79	10.63 ± 0.65	6.25 ± 0.42
		Learned perceptual image patch similarity↓	0.187 ± 0.012	0.153 ± 0.009	0.126 ± 0.008	0.114 ± 0.007	0.098 ± 0.006	0.061 ± 0.004
Places2	Irregular 50%	Peak signal-to-noise ratio↑	20.75 ± 0.41	21.93 ± 0.36	22.86 ± 0.31	23.14 ± 0.27	23.67 ± 0.23	25.32 ± 0.19
		Structural similarity index↑	0.736 ± 0.018	0.772 ± 0.015	0.801 ± 0.013	0.815 ± 0.011	0.834 ± 0.009	0.886 ± 0.006
		Fréchet inception distance↓	34.62 ± 2.15	28.37 ± 1.84	24.16 ± 1.53	21.74 ± 1.31	18.92 ± 1.07	11.36 ± 0.68
Paris StreetView	Rectangular 40%	Peak signal-to-noise ratio↑	21.26 ± 0.38	22.41 ± 0.33	23.27 ± 0.29	23.58 ± 0.25	24.13 ± 0.21	26.04 ± 0.17
		Learned perceptual image patch similarity↓	0.214 ± 0.014	0.179 ± 0.011	0.147 ± 0.009	0.132 ± 0.008	0.113 ± 0.007	0.072 ± 0.004

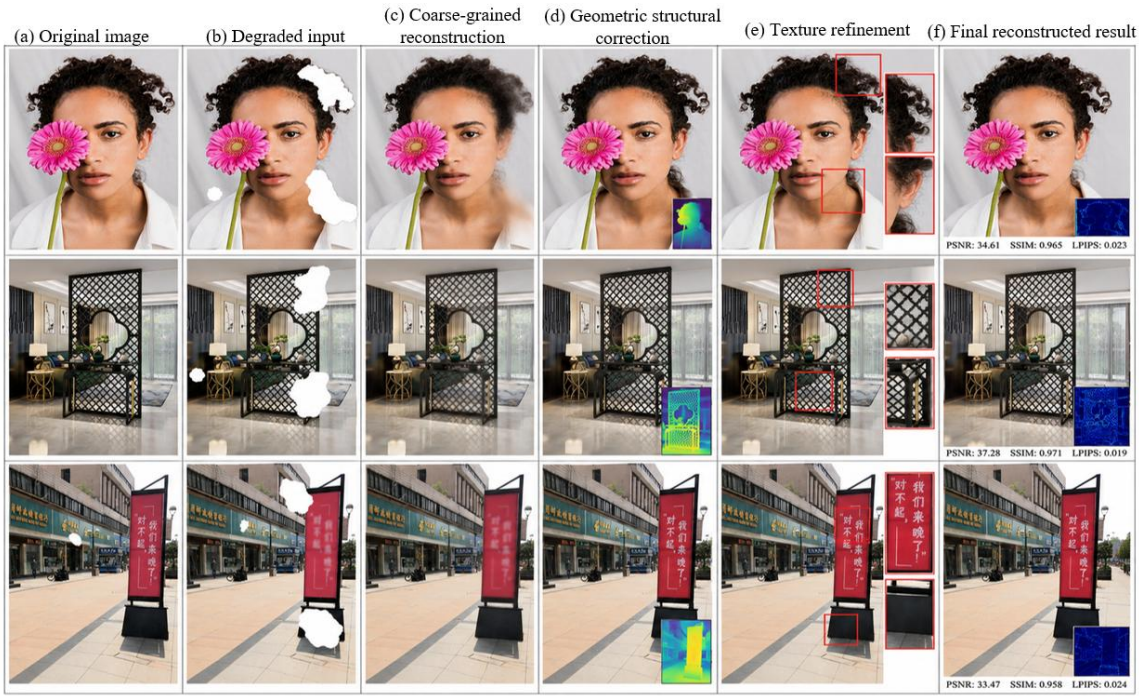

Figure 4. Structure-texture dual-branch collaborative generation process for image reconstruction and visual information completion

3.3 Ablation experiments on individual core modules

For the ablation experiments, the Places2 dataset was fixed as the evaluation benchmark, with irregular 40% missing-area masks employed for testing. Six network variants were constructed, and the performance degradation magnitude was compared by removing each core module individually. The quantitative results are presented in Table 2. Variant 0 denotes the complete structure-texture dual-branch collaborative generation network, while Variants A through E sequentially remove the context-adaptive discrepancy dynamic sampling mechanism, the lightweight geometric estimation module, the hybrid sparse attention mechanism, the multi-scale progressive reconstruction strategy, and the energy-gated feature alignment module. The metric variation Δ relative to the complete model is also recorded in the table, where decreases in peak signal-to-noise ratio and structural similarity index are denoted as negative values, and increases in Fréchet inception distance and LPIPS are denoted as positive values.

Table 2. Quantitative results of module ablation experiments and performance degradation deltas

Model Variant	Peak Signal-to-Noise Ratio ↑	ΔPeak Signal-to-Noise Ratio	Structural Similarity Index↑	ΔStructural Similarity Index	Fréchet Inception Distance↓	ΔFréchet Inception Distance	Learned Perceptual Image Patch Similarity↓	ΔLearned Perceptual Image Patch Similarity
Variant 0 (complete structure-texture dual-branch collaborative generation network)	25.17	-	0.881	-	11.72	-	0.065	-
Variant A (without the context-adaptive discrepancy)	24.06	-1.11	0.853	-0.028	15.47	+3.75	0.094	+0.029
Variant B (without the lightweight geometric estimation module)	24.32	-0.85	0.862	-0.019	14.13	+2.41	0.086	+0.021
Variant C (without the hybrid sparse attention)	24.48	-0.69	0.867	-0.014	13.46	+1.74	0.081	+0.016
Variant D (without the progressive reconstruction)	23.91	-1.26	0.847	-0.034	16.25	+4.53	0.099	+0.034
Variant E (without the energy-gated feature alignment module)	24.63	-0.54	0.871	-0.01	12.84	+1.12	0.076	+0.011

From the performance degradation deltas, the contribution of each module to the reconstruction quality can be assessed. The largest performance drop is observed when the multi-scale progressive reconstruction strategy is removed, with a peak signal-to-noise ratio degradation of 1.26 and a Fréchet inception distance increase of 4.53, demonstrating that the coarse-to-fine hierarchical generation mechanism serves as the core pillar for balancing global layout and local details. The second-largest performance loss is incurred by the removal of context-adaptive discrepancy dynamic sampling, as this module enforces contextual constraints throughout the diffusion process; in its absence, the generated regions become disconnected from the distribution of the original image, resulting in a marked increase in perceptual error. The lightweight geometric estimation module plays a critical role in ensuring three-dimensional spatial consistency; its removal leads to distortions in perspective and occlusion logic, with a pronounced deterioration observed in the Fréchet inception distance metric. The hybrid sparse attention and the energy-gated feature alignment module are primarily responsible for suppressing feature noise and reducing computational redundancy; the performance degradation incurred by their individual removal is relatively smaller, yet their combined effect effectively enhances feature representation quality in high-resolution scenarios.

Visualization heatmaps of edge sharpness and texture fidelity were simultaneously generated for the experiments. The heatmap of the complete model exhibited high-response regions uniformly covering object edges and fine textures, whereas localized response deficiencies were observed upon the removal of any module, consistent with the degradation patterns reflected in the quantitative metrics.

3.4 Robustness tests under varying missing ratios and mask shapes

In this experiment, five sets of irregular masks and three sets of rectangular masks were respectively employed, with missing-area ratios spanning the extreme range from 10% to 80%. The peak signal-to-noise ratio, Fréchet inception distance, and the proportion of failure samples were statistically computed under each condition, where failure samples were defined as image samples with learned perceptual image patch similarity values exceeding 0.35. The results are presented in Figures 5 and 6.

From the data plotted in the two figures, the trend curves of evaluation metrics with respect to missing-area ratio exhibit a monotonic degradation as the missing area increases; however, the slopes of the degradation curves show distinct piecewise characteristics. In the missing-area range of 10% to 50%, a gradual performance decline is observed. When the missing area increases to the extreme range of 70% to 80%, the performance of the baseline algorithms drops sharply, whereas the degradation curve of the proposed method exhibits a significantly lower slope. Under the 80% large-area irregular mask scenario, the proportion of failure samples for the proposed method is only 6.75%, while that of BrushNet under the same conditions exceeds 15%, validating the strong stability of the multi-scale progressive reconstruction strategy under large-area information loss scenarios. The overall

reconstruction difficulty under rectangular masks is lower than that under irregular masks with the same missing ratio, with all metrics outperforming those under equivalent-area irregular masks. The model maintains stable reconstruction capability under both mask types, demonstrating robust adaptability to varying mask shapes.

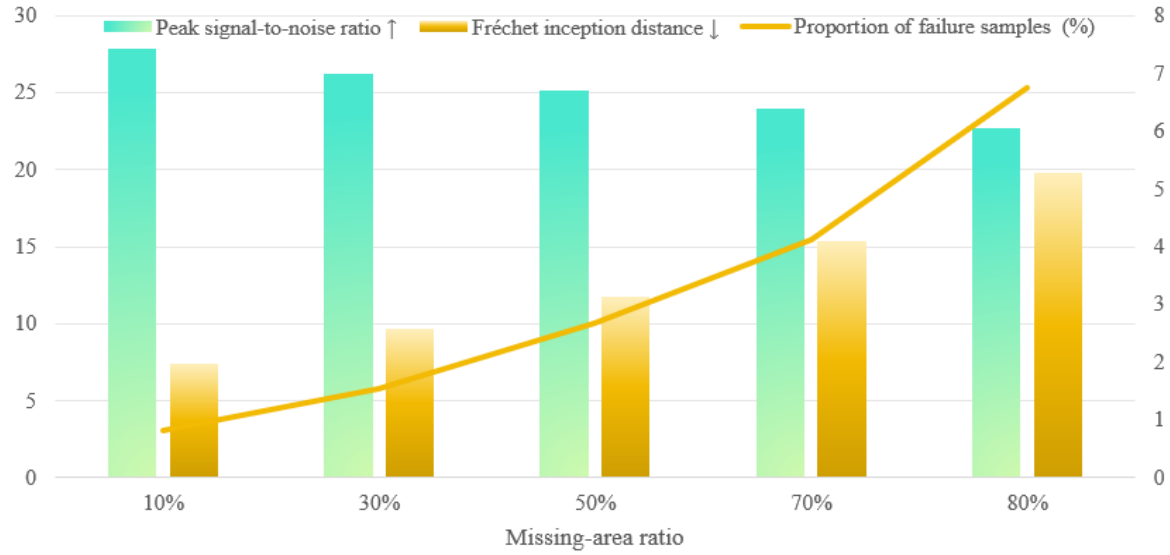

Figure 5. Model robustness metrics under irregular masks with varying missing-area ratios

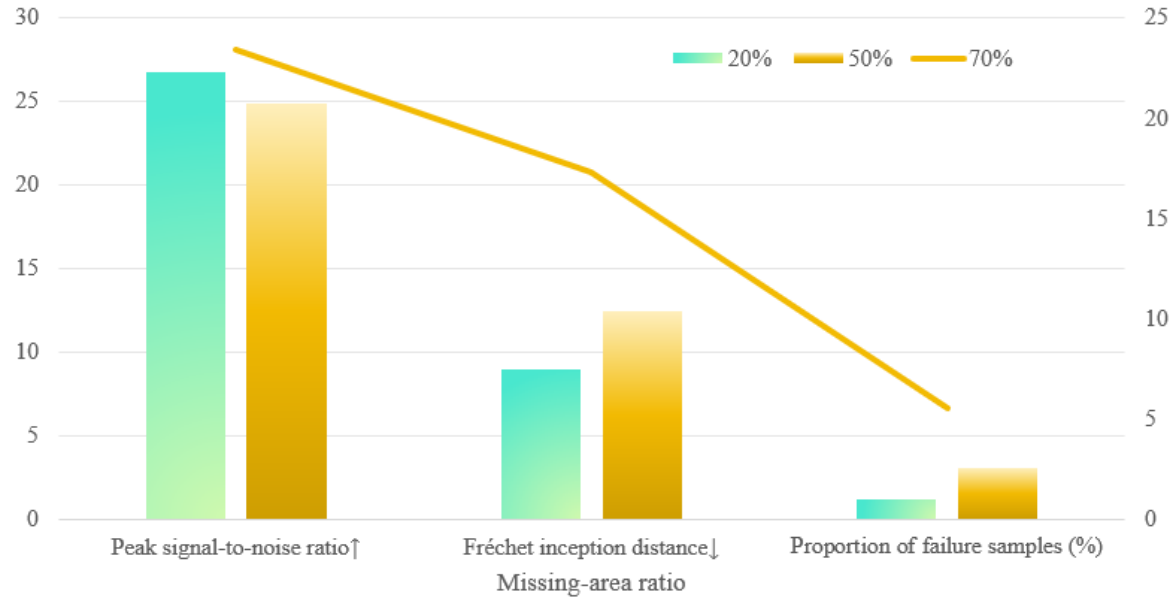

Figure 6. Model robustness metrics under rectangular masks with varying missing-area ratios

3.5 Cross-dataset generalization capability and zero-shot transfer testing

In this experiment, a leave-one-dataset zero-shot testing scheme was adopted. The model was trained jointly only on Places2 and Paris StreetView, and inference was performed directly on the face dataset CelebA-HQ and the texture dataset DTD without fine-tuning. LaMa and RePaint, which are known for their strong generalization capability, were selected as comparison baselines. The Fréchet inception distance metric and the mean opinion score averaged over 50 subjects were quantified, with the results presented in Figure 7.

From the zero-shot experimental data, it can be observed that under cross-domain scenarios, the proposed method achieves significantly lower Fréchet inception distance values than the two comparison models, with a clear advantage in subjective mean opinion scores. LaMa relies on fixed scene texture priors, making it highly susceptible to texture disarray when applied to face and pure-texture datasets. RePaint, despite leveraging diffusion-based generation, lacks general three-dimensional geometric constraints, rendering the spatial structure of cross-domain objects prone to distortion. In contrast, the proposed lightweight geometric estimation module extracts general depth and normal geometric priors in a self-supervised manner, without being tied to specific scene texture distributions. The geometric constraints derived from this module are inherently cross-scene generalizable, enabling the maintenance of reasonable spatial structure and natural texture distribution even on unseen face and texture datasets. The generalization and transfer performance of the proposed method thus substantially surpasses that of existing approaches.

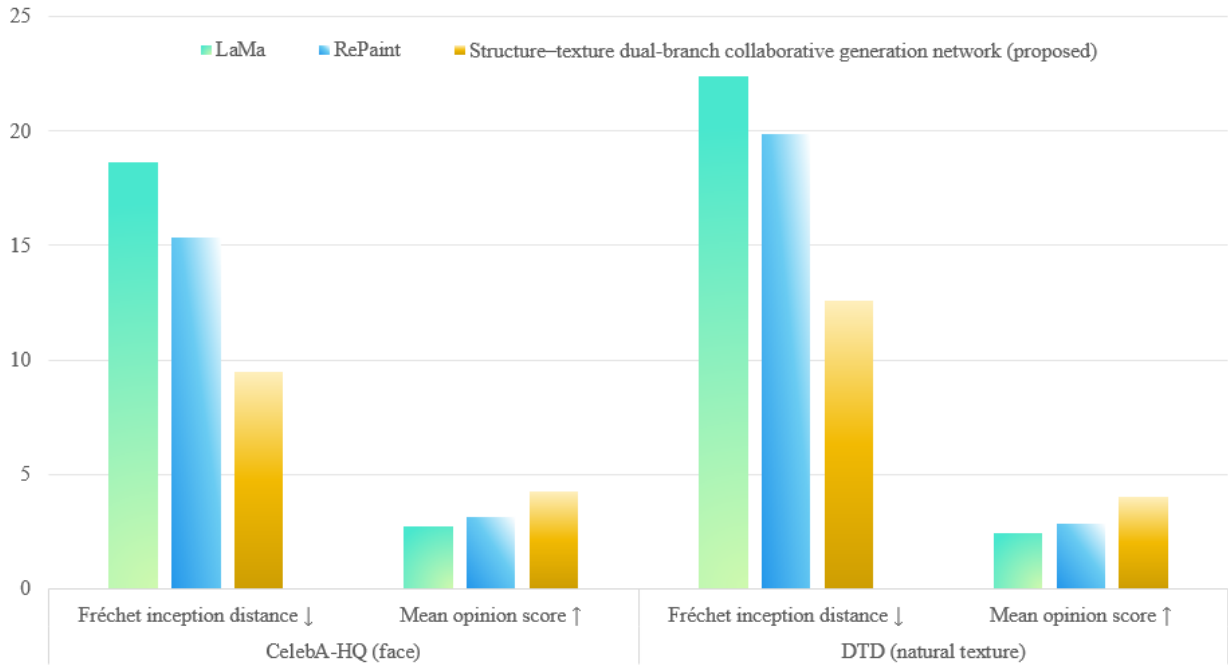

Figure 7. Zero-shot cross-dataset generalization performance comparison

3.6 Computational efficiency, parameter count, and hyperparameter sensitivity analysis

Quantitative analyses were conducted from three perspectives: model inference overhead, sparse attention hyperparameters, and context-adaptive discrepancy dynamic sampling hyperparameters. The corresponding quantitative results are presented in Table 3, Figure 8, and Table 4.

(a) Model Computational Efficiency Analysis

From in Table 3, it can be observed that the structure-texture dual-branch collaborative generation network achieves lower parameter count, floating-point operations, and single-image inference memory consumption compared to both Palette and BrushNet. The inference frame rate at 512 × 512 resolution reaches 4.82 frames per second, which is approximately 1.8 times that of BrushNet. Under 1024 × 1024 high-resolution input, the inference speed advantage is further amplified, with memory consumption being only about 60% of that of the comparison methods. The hybrid sparse attention mechanism reduces the complexity of global attention to near-linear levels, substantially reducing the inference overhead for high-resolution images while balancing reconstruction accuracy with engineering deployment efficiency.

(b) Sparsity Hyperparameter Trade-off Analysis

The data presented in Figure 8 are used to plot the Pareto curves of Fréchet inception distance versus floating-point operations. When the K/N ratio is increased from 1 to 1/4, the computational cost decreases by 45%, while the Fréchet inception distance increases by only 0.54. When the ratio is further reduced to 1/8, computational cost continues to decrease, but perceptual quality exhibits a noticeable degradation. The K/N ratio of 1/4 achieves the optimal balance between reconstruction accuracy and computational overhead, and is therefore adopted as the final sparse selection ratio in this study.

(c) Context-Adaptive Discrepancy Dynamic Sampling Hyperparameter Stability Analysis

Table 4 records the local refinement-triggered pixel ratios and perceptual errors corresponding to different scaling coefficients γ and threshold parameters β. With $\gamma = 10$ and $\beta = 0.5$ fixed, the globally optimal learned perceptual image patch similarity value is achieved. Increasing γ raises the proportion of refined pixels in edge regions, marginally reducing perceptual error, but concurrently increasing inference computational cost. Raising the threshold β reduces the refinement iteration regions, accelerating inference speed, yet degrades the continuity of defect boundary integration. Across a wide range of hyperparameter settings, the learned perceptual image patch similarity fluctuation range does not exceed 0.02, demonstrating that the context-adaptive discrepancy dynamic sampling mechanism exhibits favorable adaptive stability to hyperparameter variations, with a lower tuning threshold.

Table 3. Comparison of parameter count, computational cost, and inference efficiency across models

Model	Parameters (M)	512 × 512 Floating-Point Operations (G)	512 × 512 Frames per Second	512 × 512 Memory (GB)	1024 × 1024 Frames per Second	1024 × 1024 Memory (GB)
Palette	432.6	1284.7	2.13	14.7	0.46	27.3
BrushNet	387.2	1062.3	2.67	12.3	0.61	23.6
Structure-texture dual-branch collaborative generation network	354.1	746.2	4.82	8.6	1.34	15.2

Table 4. Sensitivity test results of context-adaptive discrepancy hyperparameters γ and β

γ	β	Refinement-Triggered Pixel Ratio (%)	Learned Perceptual Image Patch Similarity↓
5	0.3	18.4	0.083
5	0.5	12.7	0.076
5	0.7	7.2	0.071
10	0.5	16.3	0.065
20	0.5	22.6	0.063

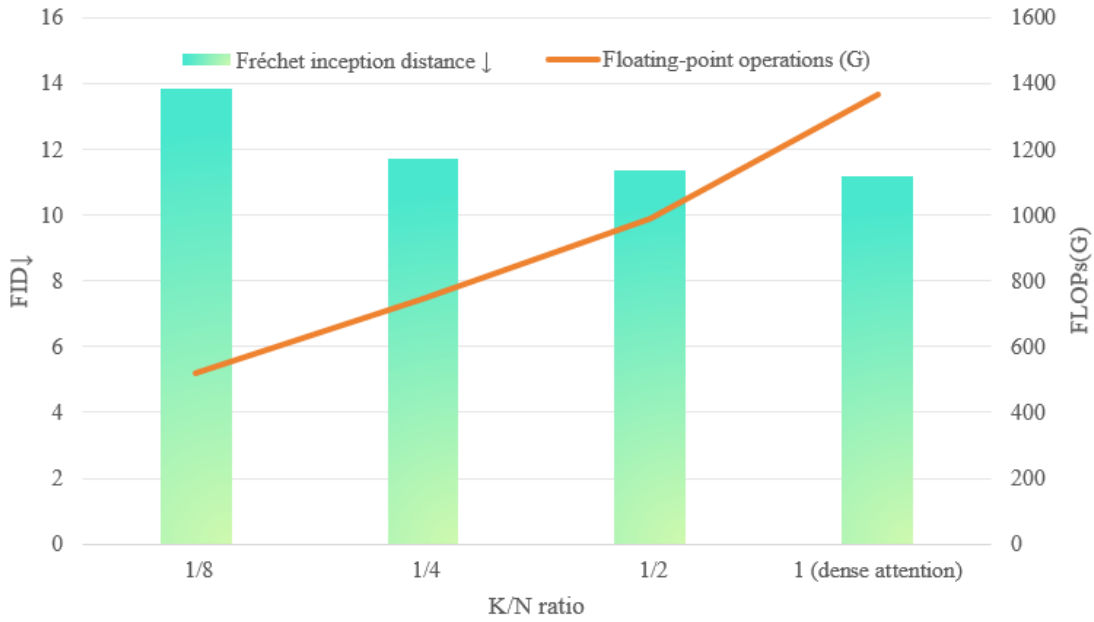

Figure 8. Performance and computational cost trade-off under different Top-K ratios in the hybrid sparse attention

4. CONCLUSION

To address the four core limitations of existing generative image reconstruction algorithms—namely, contextual discontinuity throughout the diffusion process, the absence of three-dimensional geometric constraints, computational redundancy and feature noise in attention mechanisms, and the difficulty of jointly optimizing global semantics with local textures, a structure-texture dual-branch collaborative generation network was constructed in this study. The method was built upon a latent diffusion model with a decoupled dual-branch encoding-decoding architecture, through which adaptive collaborative representation of structural contours and surface textures was achieved via spatial gated fusion. Contextual consistency constraints were embedded into the complete reverse denoising pipeline through context-adaptive discrepancy dynamic sampling, eliminating the visual discontinuities between missing regions and the original image. A lightweight self-supervised geometric estimation module was introduced, through which depth and normal priors were injected without annotations, with epipolar geometric loss serving as the constraint, ensuring that the perspective and occlusion relationships of the reconstructed content conform to physical laws. A hybrid sparse self-attention and energy-gated feature alignment module was designed, through which attention computational complexity was reduced while feature distortion induced by masks was suppressed. A multi-scale progressive reconstruction strategy was incorporated, through which global layout modeling and high-frequency texture refinement were accomplished hierarchically, reconciling the inherent conflict in multi-scale information modeling. Extensive experiments conducted on CelebA-HQ, Places2, and Paris StreetView demonstrated that the proposed method consistently outperformed multiple representative comparison algorithms across four mainstream evaluation metrics: peak signal-to-noise ratio, structural similarity index, Fréchet inception distance, and learned perceptual image patch similarity. Ablation studies validated that each sub-module contributed positively to reconstruction performance. Robustness tests under extreme degradation scenarios and zero-shot cross-dataset transfer experiments confirmed the model's stable inpainting capability and general adaptability to diverse scenes. Computational efficiency comparisons substantiated that the sparse attention architecture effectively reduced inference overhead, rendering it more suitable for real-time reconstruction of high-resolution images. The overall framework simultaneously achieves enhanced visual realism, three-dimensional physical consistency, and inference efficiency.

The current study is limited to the restoration of individual static two-dimensional images and has not yet incorporated temporal or stereoscopic spatial correlation information. In future work, temporal constraint modules will be extended based on the existing framework, through which temporally consistent inter-frame image completion schemes will be designed for continuous video sequences. Simultaneously, stereoscopic geometric modeling principles will be integrated, through which the two-dimensional inpainting network will be extended to three-dimensional scene completion tasks, and the multi-view stereo geometric constraint mechanism will be refined. The application capability of the proposed algorithm will be further extended to complex real-world scenarios,

including remote sensing time-series imagery, dynamic medical imaging, and autonomous driving perceptual systems.

REFERENCES

[1] Elharrouss, O., Almaadeed, N., Al-Maadeed, S., Akbari, Y. (2020). Image inpainting: A review. Neural Processing Letters, 51(2): 2007-2028. https://doi.org/10.1007/s11063-019-10163-0

[2] Jam, J., Kendrick, C., Walker, K., Drouard, V., Hsu, J.G.S., Yap, M.H. (2021). A comprehensive review of past and present image inpainting methods. Computer Vision and Image Understanding, 203: 103147. https://doi.org/10.1016/j.cviu.2020.103147

[3] Jiang, B., Li, X., Chong, H., et al. (2022). A deep-learning reconstruction method for remote sensing images with large thick cloud cover. International Journal of Applied Earth Observation and Geoinformation, 115: 103079. https://doi.org/10.1016/j.jag.2022.103079

[4] Czerkawski, M., Upadhyay, P., Davison, C., et al. (2022). Deep internal learning for inpainting of cloud-affected regions in satellite imagery. Remote Sensing, 14(6): 1342. https://doi.org/10.3390/rs14061342

[5] Wang, Q., Chen, Y., Zhang, N., Gu, Y. (2021). Medical image inpainting with edge and structure priors. Measurement, 185: 110027. https://doi.org/10.1016/j.measurement.2021.110027

[6] Kim, K.D., Cho, K., Kim, M., et al. (2022). Enhancing deep learning-based classifiers with inpainting anatomical side markers for multi-center trials. Computer Methods and Programs in Biomedicine, 220: 106705. https://doi.org/10.1016/j.cmpb.2022.106705

[7] Wan, Z., Zhang, B., Chen, D., et al. (2023). Old photo restoration via deep latent space translation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(2): 2071-2087. https://doi.org/10.48550/arXiv.2009.07047

[8] Wang, Z., Chen, J., Hoi, S.C.H. (2021). Deep learning for image super-resolution: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(10): 3365-3387. https://doi.org/10.1109/TPAMI.2020.2982166

[9] Wen, L.H., Jo, K.H. (2022). Deep learning-based perception systems for autonomous driving: A comprehensive survey. Neurocomputing, 489: 255-270. https://doi.org/10.1016/j.neucom.2021.08.155

[10] Muhammad, K., Hussain, T., Ullah, H., et al. (2022). Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks. IEEE Transactions on Intelligent Transportation Systems, 23(12): 22694-22715.

[11] Qin, Z., Zeng, Q., Zong, Y., Xu, F. (2021). Image inpainting based on deep learning: A review. Displays, 69: 102028. https://doi.org/10.1016/j.displa.2021.102028

[12] Zhang, X., Zhai, D., Li, T., Zhou, Y., Lin, Y. (2023). Image inpainting based on deep learning: A review. Information Fusion, 90: 74-94. https://doi.org/10.1016/j.inffus.2022.08.033

[13] Croitoru, F.A., Hondru, V., Ionescu, R.T., Shah, M. (2023). Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(9): 10850-10869. https://doi.org/10.1109/TPAMI.2023.3261988

[14] Cao, H., Tan, C., Gao, Z., et al. (2024). A survey on generative diffusion models. IEEE Transactions on Knowledge and Data Engineering, 36(7): 2814-2830. https://doi.org/10.1109/TKDE.2024.3361474

[15] Quan, W., Chen, J., Liu, Y., et al. (2024). Deep learning-based image and video inpainting: A survey. International Journal of Computer Vision, 132(7): 2367-2400. https://doi.org/10.1007/s11263-023-01977-6

[16] Elharrouss, O., Damseh, R., Belkacem, A.N., et al. (2025). Transformer-based image and video inpainting: Current challenges and future directions. Artificial Intelligence Review, 58(4): 124. https://doi.org/10.1007/s10462-024-11075-9

[17] Xiang, H., Zou, Q., Nawaz, M.A., et al. (2023). Deep learning for image inpainting: A survey. Pattern Recognition, 134: 109046. https://doi.org/10.1016/j.patcog.2022.109046

[18] Huang, W., Deng, Y., Hui, S., et al. (2024). Sparse self-attention transformer for image inpainting. Pattern Recognition, 145: 109897.

[19] Shamsolmoali, P., Zareapoor, M., Zhou, H., et al. (2025). From missing pieces to masterpieces: Image completion with context-adaptive diffusion. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(7): 6073-6087. https://doi.org/10.1109/TPAMI.2025.3558092

[20] Zhang, C., Yang, W., Li, X., Han, H. (2024). MMGInpainting: Multi-modality guided image inpainting based on diffusion models. IEEE Transactions on Multimedia, 26: 8811-8823. https://doi.org/10.1109/TMM.2024.3382484

[21] Cao, C., Dong, Q., Fu, Y. (2023). ZITS++: Image inpainting by improving the incremental transformer on structural priors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(10): 12667-12684. https://doi.org/10.1109/TPAMI.2023.3280222

[22] Khan, M.A.U., Nazir, D., Pagani, A., et al. (2022). A comprehensive survey of depth completion approaches. Sensors, 22(18): 6969. https://doi.org/10.3390/s22186969

[23] Han, K., Wang, Y., Chen, H., et al. (2023). A survey on vision transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(1): 87-110. https://doi.org/10.1109/TPAMI.2022.3152247

[24] Tay, Y., Dehghani, M., Bahri, D., Metzler, D. (2023). Efficient transformers: A survey. ACM Computing Surveys, 55(6): 109. https://doi.org/10.1145/3530811

[25] Zeng, Y., Fu, J., Chao, H., Guo, B. (2023). Aggregated contextual transformations for high-resolution image inpainting. IEEE Transactions on Visualization and Computer Graphics, 29(7): 3266-3280. https://doi.org/10.48550/arXiv.2104.01431