applied sciences

MDPI

Article

Structure-Guided Image Inpainting Based on Multi-Scale Attention Pyramid Network

Jun Gong [1,†], Senlin Luo [1,†], Wenxin Yu [2] and Liang Nie [2,*]

[1] Information System and Security & Countermeasures Experimental Center, Beijing Institute of Technology Beijing, Beijing 100081, China

[2] School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang 621010, China; yuwenxin@swust.edu.cn

[*] Correspondence: jianglihuakai@gmail.com

[†] These authors contributed equally to this work.

Abstract: Current single-view image inpainting methods often suffer from low image information utilization and suboptimal repair outcomes. To address these challenges, this paper introduces a novel image inpainting framework that leverages a structure-guided multi-scale attention pyramid network. This network consists of a structural repair network and a multi-scale attention pyramid semantic repair network. The structural repair component utilizes a dual-branch U-Net network for robust structure prediction under strong constraints. The predicted structural view then serves as auxiliary information for the semantic repair network. This latter network exploits the pyramid structure to extract multi-scale features of the image, which are further refined through an attention feature fusion module. Additionally, a separable gated convolution strategy is employed during feature extraction to minimize the impact of invalid information from missing areas, thereby enhancing the restoration quality. Experiments conducted on standard datasets such as Paris Street View and CelebA demonstrate the superiority of our approach over existing methods through quantitative and qualitative comparisons. Further ablation studies, by incrementally integrating proposed mechanisms into a baseline model, substantiate the effectiveness of our multi-view restoration strategy, separable gated convolution, and multi-scale attention feature fusion.

Keywords: image inpainting; image processing; multi-scale attention pyramid; multi-view inpainting

check for updates

Citation: Gong, J.; Luo, S.; Yu, W.; Nie, L. Structure-Guided Image Inpainting Based on Multi-Scale Attention Pyramid Network. *Appl. Sci.* **2024**, *14*, 8325. https://doi.org/10.3390/app14188325

Academic Editor: Pedro Couto

Received: 30 July 2024
Revised: 1 September 2024
Accepted: 8 September 2024
Published: 15 September 2024

1. Introduction

Image inpainting [1–4] technology utilizes the known information of an image to fill in the missing areas, serving as a crucial branch in the field of computer vision. The outcomes produced by image inpainting algorithms are ideally consistent with the original in terms of texture, structure, color style, and semantics. This technique finds widespread application in various areas, including image editing [5–7], cultural heritage preservation [8], old photo restoration [9], image super-resolution [10,11], and object removal [12,13]. Image inpainting has evolved significantly over the decades, during which a plethora of remarkable methods for image inpainting has emerged. The techniques from different eras bear distinct characteristics of their times. Initially, image inpainting methods could broadly be categorized into two types: patch-based and diffusion-based methods [14–17]. The fundamental principle of patch-based restoration involves identifying small segments within the known areas of an image that closely match the texture, color, and structure of the missing regions. These segments, or "patches", are then copied and pasted into the voids, thereby achieving the restoration and integrity of the image. On the other hand, diffusion-based image inpainting methods, also known as PDE (partial differential equation)-based methods [17,18], rely on mathematical diffusion processes to seamlessly fill in the gaps or damaged areas of an image. This approach utilizes information from the image's known regions, employing mathematical models to infer the possible content of the missing areas. Both categories struggle to generate semantic content and, while they can effectively repair small areas of missing data, they falter with larger gaps, hence being referred to as traditional image inpainting methods.

With the advancement of societal technologies, the emergence of deep learning has propelled developments across various domains, including object recognition [19], object tracking [20], target classification [21], and image inpainting. The pioneering use of deep learning for image inpainting was by Pathak and colleagues [22], who leveraged the capability of deep learning technologies to generate new semantic content from known content to fill in missing areas of images. Subsequently, a plethora of image inpainting methods based on deep learning [22–27] emerged, predominantly utilizing a single degraded image as input. This approach benefits from lower computational resource consumption but struggles to acquire sufficient effective information, especially in cases of images with large missing areas. This leads to issues such as inconsistent textures, structural distortions, and artifacts in the inpainting results.

In order to enable networks to extract and learn more comprehensive image features, an increasing number of image inpainting methods have begun to employ multi-view inpainting [28]. These methods generate auxiliary views—such as edge views [29], contour views [30], and views at different resolutions [31]—from the original image to aid in the inpainting process, providing additional information to enhance the inpainting outcomes. However, the use of multi-view inpainting introduces the risk of propagating errors from the auxiliary views into the final inpainting result, as incorrect content from other views can be incorporated during the inpainting process. Thus, ensuring the accuracy and reliability of the generated auxiliary views becomes crucial. Moreover, as the use of multi-view aids increases computational costs, efficiently leveraging these auxiliary views also becomes critically important.

In addition to generating content for missing areas in images, a successful image inpainting result must also maintain stylistic and visual consistency and coherence between the generated content and the existing areas. Attention mechanisms [32–34] provide an effective means to ensure the consistency of image inpainting content by focusing on key parts of the image and efficiently utilizing contextual information, significantly improving the quality of image inpainting. However, this approach only guarantees contextual consistency between the generated content and the existing content, failing to ensure internal consistency within the generated content itself, leading to parts of the generated content appearing visually discordant. To achieve visually satisfying repair effects, the internal consistency of the generated content is equally crucial.

Current popular image inpainting methods often emphasize achieving large receptive fields, sometimes at the expense of considering the importance of information near the edges of missing areas. However, in the inpainting process for irregular missing areas, not all shapes necessitate assistance from distant information. For instance, when dealing with missing areas resembling spider webs, information surrounding the missing regions is often more critical for accurate inference, while distant information may introduce noise and negatively affect the inpainting results. This phenomenon has been discussed in prior works, which highlight the challenges of irregular missing areas for traditional image-inpainting techniques (e.g., Liu et al. [26]).

Overall, the existing image inpainting methods have the following shortcomings in generating satisfactory results: (1) They only rely on a single damaged image for inference, without effectively utilizing the remaining known information within the image. (2) They fail to fully consider the internal consistency of the generated content in the missing regions. (3) During the inpainting process, they overly focus on the distant information from the missing regions and adopt the same receptive field strategy for all types of missing regions. Therefore, this paper proposes an structure-guided inpainting method based on a multi-scale attention pyramid structure. This method consists of two sub-networks: one for generating a complete edge structure and the other utilizing the generated edge structure map as auxiliary information to guide the image inpainting, ultimately producing a structurally constrained complete image. Utilizing additional edge structures as guidance addresses the issue of insufficient utilization of effective information in existing inpainting methods for missing images.

Furthermore, within the inpainting network, this paper introduces a spatial attention mechanism using a pyramid structure that can fully utilize multi-scale feature information from the missing image, enhancing the internal consistency of the generated content in the missing regions. Additionally, this cross-scale attention mechanism can expand the receptive field for local information, and the fusion of multi-scale information avoids artifacts in the inpainting results due to limited receptive fields in a single layer.

The main contributions of the proposed method can be summarized as follows: (1) Generating accurate edge structure maps as auxiliary information to guide image inpainting, achieving multi-view image inpainting, and obtaining visually satisfying results. (2) Proposing a multi-scale feature fusion mechanism that better preserves shallow texture information and deep structural information of the image, enabling the cross-scale attention mechanism to fuse features that are more beneficial for inpainting results. (3) Introducing a cross-scale attention mechanism based on a pyramid structure, where each scale captures different levels of detail in the image. Utilizing the multi-scale attention mechanism for feature weighting enhances the fineness and internal consistency of the inpainting results.

2. Related Work

After decades of development, image inpainting can be generally categorized into two main classes: traditional inpainting methods and deep learning-based inpainting methods. The former primarily consists of patch-based and diffusion-based methods. Among the latter, inpainting techniques utilizing adversarial networks have become the most popular. Patch-based methods fill the missing areas of a degraded image by searching within its known regions for patches that could closely resemble the missing area. Initially dominated by techniques such as the region filling approach proposed by Criminis [14], these methods demonstrated outstanding results at the time. However, the search for similar patches was time-consuming and computationally intensive. To address this, Barnes et al. [15] introduced the PatchMatch method, which employs a randomized nearest-neighbor matching scheme to significantly reduce the computational load and inpainting time. Diffusion-based methods employ mathematical techniques to gradually spread pixel values from the surrounding areas into the missing regions, following specific rules. Bertalmio et al. [16] were pioneers in applying diffusion techniques in image inpainting, adopting a rule that fills inwards from the outer boundary, mimicking the manual restoration of artworks. Chan and Shen [17] later proposed a curvature-driven diffusion rule, which overcame the limitations of previous methods that were only effective in restoring densely textured images, enabling better transmission of edge information and inpainting of non-textured images. While traditional inpainting methods achieved satisfactory results on small missing areas and simple texture-style images, they fundamentally lacked an understanding of semantic information and the ability to generate new content. Thus, they were often inadequate for large missing areas, resulting in the creation of pixel blocks instead of meaningful content.

With the advent of deep learning in computer vision, traditional inpainting methods began to fade. Pathak et al. [22] were among the first to apply deep learning to image inpainting, introducing an encoder–decoder architecture that transforms damaged images into complete ones. However, their method's adversarial loss was roughly applied to the missing areas, leading to a stark contrast between the generated content and the existing areas, and low consistency. To improve this, Iizuka et al. [35] proposed the concept of global and local discriminators to enhance the consistency between the overall image and the content generated for the missing areas. This dual-discriminator mechanism, however, was only effective for regularly shaped missing areas and performed poorly on irregular ones. Liu et al. introduced partial convolution [26], which uses an automatic mask update mechanism to reduce the interference of invalid information during inpainting. Yet, this method becomes ineffective in deeper networks as the mask gradually diminishes.

These approaches commonly used a single view for image inpainting, which, while easy to implement, often lacked sufficient constraints and reference information, leading to poor performance in complex scenes. To counter this, researchers proposed multi-view inpainting, which leverages features of the missing image from multiple perspectives to achieve comprehensive repair. Methods like EdgeConnect by Nazeri et al. [36] use edge maps extracted using the Canny edge detector as an additional view, providing more context and structural guidance for the inpainting process.

Despite the excellence of the Canny operator in edge detection, its robustness against complex structures or noise can be insufficient, potentially compromising the subsequent inpainting. The E2I (edge-to-image) method introduced by Xu et al. [37], which employs the holistically nested edge detection (HED) technique, brought unique advantages to image inpainting. Yet, the reliance on deep convolutional neural networks (CNNs) for HED necessitates extensive annotated data for training, posing challenges in complex inpainting tasks due to data demands, computational complexity, and accuracy of edge detection. Xiong et al. [30] used DeepCut technology to predict precise boundary salient object masks, or contour images, providing clearer edge information for guidance in inpainting. However, the accuracy of DeepCut's predictions can be affected by training data and model complexity, potentially leading to blurred or distorted edges in the inpainting results. Ren et al.'s structure–flow [38] method uses edge-preserving smooth images to provide global structural information during the second phase of inpainting, maintaining overall image structure but sometimes resulting in overly smoothed areas lacking realism. Texture-aware multi-GAN approaches by Hedjazi and Genc [31], involving multiple GAN networks at different resolutions, capture varied texture details but require substantial computational resources and time, and coordination between networks can be challenging, possibly leading to inconsistencies or artifacts in the inpainted images. These methods, based on multi-view approaches, depend on the accuracy of additional views, which, if erroneous, can adversely affect the final inpainting results. Therefore, ensuring the accuracy of each view and applying appropriate constraints is crucial in multi-view inpainting to prevent potential negative impacts on the results. Additionally, selecting the right method based on the specific task and data characteristics, possibly in combination with other techniques, is necessary to optimize the inpainting outcomes.

The application of attention mechanisms in the field of image inpainting has significantly improved model performance. Yu et al. [39] introduced a context attention module that can borrow or copy feature information from known background blocks to fill in the missing parts by simulating the long-term correlation between distant context information and the missing area. However, this method has its drawbacks: The generated inpainting content might be inaccurate or distorted if the distant context information is only weakly related to the missing area or if there are complex texture and structural changes. Zheng et al. [40] designed a short- and long-term context attention layer that enhances appearance consistency by leveraging the distance relationship between decoder and encoder features. Nevertheless, this approach has high computational resource demands, especially for high-resolution images, potentially resulting in longer inpainting processes. Moreover, inaccuracies in the correspondence between encoder and decoder features can lead to biased inpainting outcomes. Zeng et al. [41] proposed the PEN-Net, utilizing a cross-layer attention transfer mechanism to ensure visual and semantic consistency by learning regional affinity from high-level feature maps to guide the repair of adjacent lower layers. However, this method demands a high level of model complexity, necessitating careful design and adjustment of the network structure. Furthermore, if the regional affinity learning in high-level feature maps is inaccurate, it might cause errors or inconsistencies in the inpainting results of the lower layers. While deep learning techniques improve image inpainting by deepening the network architecture and expanding the receptive field, this strategy can sometimes neglect the significance of the surrounding context. This oversight may cause problems such as gradient vanishing or explosion, thus destabilizing the training process. Moreover, while increasing the receptive field size helps to capture more contextual information, it may also bring in unwanted noise and interference. Consequently, this can lead to pronounced color discrepancies and the introduction of undesired details in the inpainting results.

3. Methodology

The methodology proposed in this document is illustrated in Figure 1. Overall, the structure-guided multi-scale attention pyramid Network for image inpainting can be divided into two main parts. The first part consists of a structure repair network with strong constraints, employing a U-Net-based encoder–decoder network as its backbone, a structure that has been extensively validated for its effectiveness. Additionally, this part includes an adversarial network with the same architecture, which uses real structures for feature extraction to constrain the structure prediction network. After the edge structure repair network generates the structural view, the second part, the cross-scale attention pyramid repair network, commences operation. The core strength of this network lies in its ability to fully utilize multi-scale information of the image and enhance the extraction and repair of key features through the attention mechanism. Initially, the cross-scale attention pyramid repair network performs multi-scale feature extraction on the input structural view. Through a pyramidal hierarchical structure, the network captures feature information from global to local scales, facilitating a comprehensive understanding of the image's structure and details.

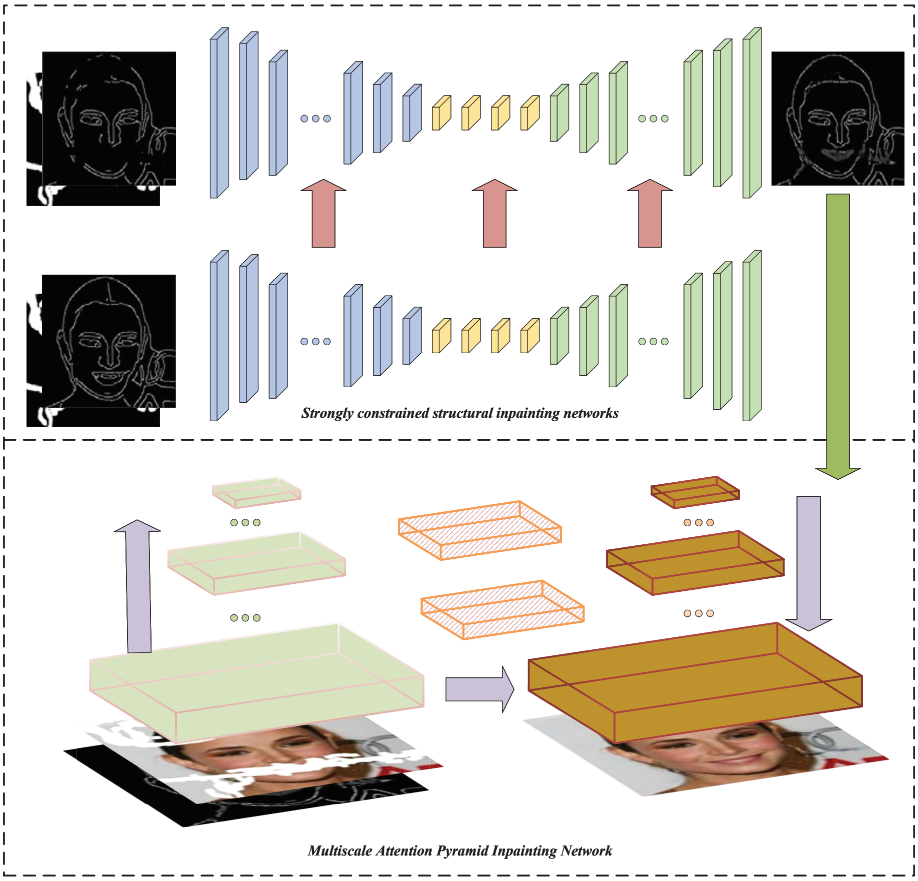
Strongly constrained structural inpainting networks
Multiscale Attention Pyramid Inpainting Network

Figure 1. Overall architecture of the proposed image inpainting network, consisting of two main components: the structural restoration network and the multi-scale attention pyramid network.

3.1. Dual U-Net Constrained Structure Prediction Network

The structure prediction of missing image regions directly influences the quality of the final image inpainting results. The detailed architecture of the proposed dual U-Net constrained structure network is depicted in Figure 2. This network serves as the first component of the entire repair system, aiming to generate edge images that closely resemble the real image structure, thereby enhancing the accuracy of the image inpainting outcomes. The network features two identical branches of generative adversarial networks (GANs), each consisting of a U-Net-based generator [42] and a PatchGAN-based discriminator [43]. The U-Net extracts features from the input image through its encoder, capturing structural and textural information across multiple scales. This multi-scale feature extraction capability enables U-Net to fully understand the context of the image, providing robust support for the subsequent repair process. In the decoder segment, U-Net gradually restores image detail by upsampling and merging with high-resolution features from the encoder, allowing the repaired image areas to seamlessly integrate with surrounding regions, maintaining the overall integrity and consistency of the image. Additionally, the skip-connection mechanism introduced by U-Net allows for direct transmission of low-level and high-level feature information, thus avoiding information loss and aiding in the recovery of more detailed image features.

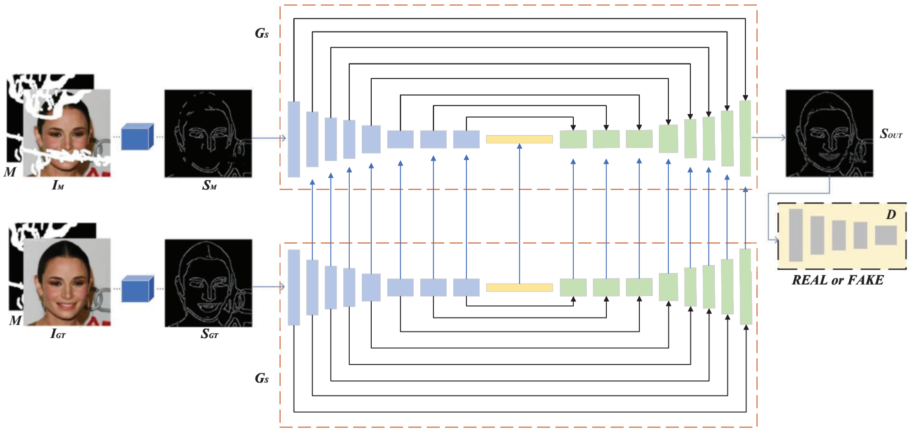

Figure 2. Specific structure of an edge-enhancement network for image inpainting under strict constraints. The network comprises two encoder–decoder networks with identical structures, differing primarily in the input images. By extracting edge features from real images, the network minimizes the discrepancy between the predicted and original structures, thereby enhancing the accuracy of the reconstruction in damaged or incomplete areas.

In the structural repair sub-network, the real image is denoted as I_{GT}, the mask as M (where 1 represents known areas, and 0 represents missing areas), and the damaged image as I_M ($I_M = I_{GT} \odot (1 - M)$).

The real structural view S_{GT} and the damaged structural view S_M are obtained through the Canny operator applied to the real and damaged images, respectively. The final output of this network is a complete structural view S_{OUT}. S_{OUT} participates in the repair tasks of the second part, the multi-scale attention pyramid network, serving as critical guiding information. The accuracy of S_{OUT} is paramount. To enhance the accuracy of the generated results, a strong constraint mechanism is proposed, employing another U-Net-based encoder–decoder network for feature extraction from the real structural view, then imposing feature constraints on each corresponding layer of the damaged structure extraction network. The discrepancy between the structural features extracted by the two branches is computed as follows:

$$L_{\text{mul}} = \sum_{i \in N} \| G_{s_i}(S_M) - G_{s_i}(S_{GT}) \|_2 \tag{1}$$

where N represents the total number of layers in G_s. G_s represents the image structure generator. By reducing this discrepancy, the generated damaged structural view is ensured to closely resemble the real structure. Moreover, reconstruction loss, adversarial loss, and feature matching loss are also employed to enhance the accuracy of the structural view prediction. The reconstruction loss, adversarial loss, and feature matching loss are defined, respectively, as

$$L_r = \|S_{\mathrm{OUT}} - S_{\mathrm{GT}}\|_2 \tag{2}$$

$$L_{\mathrm{adv}} = \min_{G_s} \max_{D_s} \mathbb{E}_{S_{\mathrm{GT}}}[\log D_s(S_{\mathrm{GT}})] + \mathbb{E}_{S_M}[\log(1 - D_s(G_s(S_M, M)))] \tag{3}$$

$$L_{\mathrm{fm}} = \mathbb{E}\left[\sum_i \frac{1}{N_i} \|D_s^{(i)}(S_{\mathrm{GT}}) - D_s^{(i)}(S_{\mathrm{OUT}})\|_1\right] \tag{4}$$

where $D_s^{(i)}$ represents the i-th layer of the discriminator D_s. The feature matching loss compares the feature discrepancies at all intermediate layers of the discriminator, thereby enhancing the accuracy of the structural view and the stability of adversarial network training.

We include all losses above as the total loss:

$$L_{total} = \lambda_{mul} L_{mul} + \lambda_r L_r + \lambda_{adv} L_{adv} + \lambda_{fm} L_{fm} \tag{5}$$

where $\lambda_{mul} = 8$, $\lambda_r = 5$, $\lambda_{adv} = 1$ and $\lambda_{fm} = 100$. The aforementioned loss weights are empirically determined through experiments. This combination of weights allows for a balanced performance in both texture and semantic restoration, ensuring that the generated images achieve high fidelity in terms of structural accuracy and perceptual quality. By carefully tuning these weights, the model is capable of producing results that maintain fine texture details while also preserving the overall semantic coherence of the image.

3.2. Multi-Scale Attention Pyramid Network

The multi-scale attention pyramid network for image inpainting, as detailed in Figure 3, operates under the guidance of previously restored structural views to complete the image repair process. This network is the core component of the image inpainting task; without it, image repair would not be feasible. Structured as a pyramid, each layer incorporates a multi-scale attention fusion module proposed in this document, which is critical for feature extraction and propagation. Feature extraction is essential for the image inpainting task. To achieve more accurate and comprehensive feature learning, we propose a separable gated convolution strategy embedded within the encoder of the pyramid network. The separable masked update convolution first applies conventional convolution to the input feature maps, generating a set of intermediate feature maps. Subsequently, a gating operation is performed on these intermediate feature maps to obtain a set of weight maps. By separating the two operations, the separable masked update convolution reduces the number of convolution kernels and parameters in the model while achieving similar or even superior performance to the original gated convolution method. Additionally, the separable masked update convolution allows for greater flexibility in model architecture design and enhances the model's ability to learn complex representations. The operating principle of the separable gating mechanism is shown in Figure 4.

According to the operating principle depicted in Figure 4, the separable masked update convolution follows several steps. First, the group parameter of the convolution kernel is set equal to the number of input channels, resulting in an output feature map count equal to the number of input channels, assuming the input channels number N. Thus, N feature maps are generated. These N feature maps are then divided into two groups at a ratio of N − 1:1. The first group is activated by the ReLU function, while the second group is activated by the sigmoid function. The rationale behind using two different activation functions is to provide diverse nonlinear transformations to the feature maps. Once activated, the two groups are multiplied to obtain weighted feature maps. Finally, the weighted feature maps pass through a convolution layer composed of filters with a kernel size of 1. This convolution layer helps to expand the output channels and generate new, deeper feature maps. By employing the separable masked update convolution method, the model can learn more complex representations with fewer parameters, resulting in better performance and faster convergence during training.

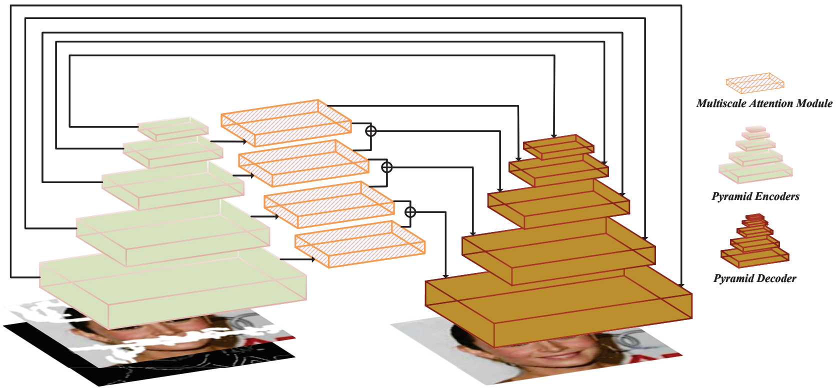

Figure 3. Structure of a multi-scale attention pyramid restoration network which includes a pyramid-structured encoder, decoder, and multi-scale attention modules. The network leverages edge maps obtained from a structural generation network to assist in generating a fully restored image.

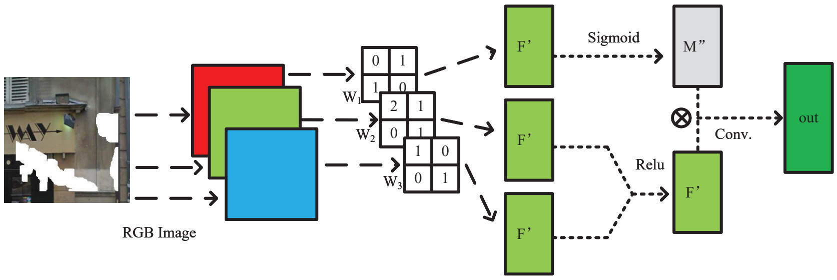

Figure 4. Steps of the separable gated convolution operation, detailing the entire convolution process from the input image to the output image.

Separable masked update convolution reduces the impact of erroneous information during the inpainting process by introducing a gating mechanism that dynamically adjusts convolutional weights based on the reliability of the input data. This mechanism selectively filters out irrelevant or incorrect information, ensuring that only valuable and accurate features are used in the reconstruction. Additionally, the separable nature of the convolution operation, which splits the convolution into depthwise and pointwise steps, provides finer control over how information is processed, reducing the risk of error accumulation. By integrating multi-scale information fusion, the method also effectively balances local and global context, further minimizing the influence of misleading data. This selective and controlled information propagation makes separable masked update convolution particularly effective at mitigating the negative effects of errors in the inpainting process, leading to more accurate and robust inpainting results.

The number of parameters required for standard convolution is as follows:

$$\begin{aligned} N_{conv} &= K_s \times K_s \times O_c \times I_c \\ &= (K_s \times K_s \times I_c) \times O_c \end{aligned} \tag{6}$$

The number of parameters required for separable masked update convolution is as follows:

$$\begin{aligned} N_{smuc} &= K_s \times K_s \times In_c + O_c \times 1 \times 1 \times I_c \\ &= (K_s \times K_s + O_c) \times I_c \\ &= (K_s \times K_s \times I_c) \times (1 + \frac{O_c}{K_s \times K_s}) \end{aligned} \tag{7}$$

N_{conv} and N_{smuc} represent the number of parameters required for standard convolution and separable masked update convolution (SMUC), respectively. K_s denotes the kernel size, while O_c and I_c represent the number of output and input channels, respectively. When $K_s = 1$, SMUC requires I_c more parameters than standard convolution. However, as the kernel size increases, SMUC requires fewer parameters compared to standard convolution. The proposed method utilizing SMUC can achieve nearly a 30 percent reduction in parameters.

To capture more precise features, once the pyramid context encoder learns latent features, it transmits them from higher to lower semantic features through the multi-scale attention fusion module. Given a pyramid context encoder composed of L layers, the feature maps from deeper to shallower layers are represented as $\phi^L, \phi^{L-1}, \ldots, \phi^1$, with each layer's features constructed by the multi-scale attention fusion operation:

$$\begin{aligned} \psi^{(L-1)} &= f(\phi^{(L-1)}, \phi^L), \\ \psi^{(L-2)} &= f(\phi^{(L-2)}, \psi^{(L-1)}), \\ &\vdots \\ \psi^1 &= f(\phi^1, \psi^2) = f(\phi^1, f(\phi^2, \ldots, f(\phi^{(L-1)}, \phi^L))). \end{aligned} \tag{8}$$

The function f represents the multi-scale attention fusion operation. Attention is typically obtained by the affinity between blocks inside and outside a region (usually 3 × 3), allowing relevant features to be transferred (i.e., weighted replication through affinity in the context) into the internal region. The output of each fusion operation is denoted as ψ^l, which is then used as an input for the next layer's fusion process. The multi-scale attention fusion module initially learns regional affinity from the high-level feature map ψ^l. It extracts blocks from ψ^l and calculates the cosine similarity between blocks inside and outside the missing area:

$$s^l_{(i,j)} = \left(\frac{p^l_i}{\|p^l_i\|_2}, \frac{p^l_j}{\|p^l_j\|_2} \right) \tag{9}$$

where p^l_i is the i-th block extracted from outside the mask in ψ^l, and p^l_j is the j-th block. A softmax is then applied to obtain the attention scores for each block based on the similarity:

$$\alpha^l_{(j,i)} = \frac{\exp(s^l_{(i,j)})}{\sum_{i=1}^{N} \exp(s^l_{(i,j)})} \tag{10}$$

After obtaining the attention scores for the high-level feature map, the gaps in its adjacent lower-level feature map can be filled:

$$p^{(l-1)}_j = \sum_{i=1}^{N} \alpha^l_{(j,i)} p^{(l-1)}_i \tag{11}$$

where $p^{(l-1)}_i$ is the i-th block extracted from outside the mask in $\phi^{(L-1)}$, and $p^{(l-1)}_j$ is the j-th block to be filled in the missing region. After computing all blocks, we can ultimately achieve feature filling by transferring attention from ψ^l.

4. Experimental Results and Datasets

4.1. Datasets

The experimental evaluation of this study was conducted on two widely recognized image datasets, each focusing on distinct subject matters: urban street scenes and celebrity faces. The Paris Street View dataset comprises high-resolution images captured from various street perspectives in Paris, providing a rich variety of architectural details and urban elements. Meanwhile, the CelebA dataset is a large-scale collection of celebrity faces, containing over 200,000 annotated images. This dataset is extensively used in facial recognition and image restoration research due to its diversity in facial expressions, poses, and lighting conditions. The application of these datasets ensures a comprehensive evaluation of the model's performance across different types of visual data, thus enhancing the diversity of experiments and ensuring the generalizability of the results.

To accurately simulate various types of damage within images and evaluate the model's ability to restore them under different conditions, this study employs a mask dataset introduced by Liu et al. [26] This dataset includes 12,000 images with various extents of damage, categorized into six levels based on the proportion of missing regions, ranging from 0% to 60% at increments of 10%. The masks are applied to simulate real-world image degradation scenarios, allowing the model to learn how to handle and restore images with different levels of occlusion or damage.

The neural network is implemented using the PyTorch (version 1.10.1, developed by Facebook's AI Research lab (FAIR), sourced from Menlo Park, CA, USA) framework, a popular deep learning library known for its flexibility and efficiency. The model is trained and optimized using the Adam optimizer, with a learning rate initially set to 1×10^{-4} and beta values $\beta_1 = 0.5$ and $\beta_2 = 0.9$. The network is trained for 200 epochs with a batch size of 16, ensuring adequate learning and convergence. To prevent overfitting, a learning rate decay schedule is employed, reducing the learning rate by a factor of 0.1 every 50 epochs. The combination of these parameters is crucial for achieving stable and efficient training, leading to a model that is capable of restoring images with high fidelity in terms of both texture and semantic content.

4.2. Quantitative Comparison

To thoroughly assess the effectiveness of the proposed image inpainting framework, this study utilizes two commonly employed metrics in the field of image inpainting: pixel similarity and structural similarity index measure (SSIM). These metrics are pivotal for evaluating the accuracy with which the inpainted images replicate the pixel-level details and maintain structural integrity compared to the original images. The mean absolute error (MAE) is a widely utilized metric in the field of image restoration and repair, offering a straightforward way to evaluate the performance of models that aim to reconstruct or inpaint damaged, corrupted, or missing parts of an image. MAE calculates the average of the absolute differences between the predicted pixel values of the restored image and the corresponding pixel values of the ground truth (original, undamaged) image. In addition, the Fréchet inception distance (FID) is incorporated as an essential supplementary metric to gauge the visual perceptual quality of the restored images. FID is particularly useful in capturing the holistic image quality and the fidelity of textures and colors as perceived by the human visual system. The experimental setup includes rigorous comparative tests against several cutting-edge methods in the field, specifically FFTI [44], SWT [45], FMGDN [46], and MFMAM [3]. These methods represent a diverse set of approaches, ranging from texture-focused techniques to those prioritizing structural integrity in image inpainting. The results of these comparisons are meticulously documented in Tables 1 and 2, where the superior performance of the proposed method is clearly delineated across all the metrics used. The analysis includes tests on two well-known datasets: Paris Street View and CelebA. The Paris Street View dataset, known for its urban landscapes and architectural elements, tests the method's ability to handle images with straight lines and geometric structures. The CelebA dataset, on the other hand, evaluates how well the method restores human facial features, which require precise treatment of organic shapes and skin textures. The study's findings underscore that the proposed method not only achieves higher pixel and structural similarity scores but also significantly improves the visual perceptual quality of the restored images, as evidenced by its lower FID scores compared to the benchmarks. This enhancement in performance can be attributed to the innovative integration of multi-scale attention mechanisms and the use of structure-guided strategies, which together facilitate a more nuanced and context-aware restoration process.

Table 1. These are the experimental data of the proposed image inpainting method in this paper with several other image inpainting methods on the Paris Street View dataset, where the second column represents the percentage of damaged area in the damaged image.

Metrics		**Methods**				
		FFTI [44]	**SWT [45]**	**FMGDN [46]**	**MFMAM [3]**	**Proposed**
PSNR↑	(0.1,0.2]	29.1022	29.6467	30.0476	29.9150	**30.9707**
	(0.2,0.3]	26.0733	26.1348	26.7547	27.2925	**27.5798**
	(0.3,0.4]	24.6046	24.2154	24.2957	25.4415	**25.4962**
	(0.4,0.5]	23.1416	22.6202	22.5091	**24.0152**	23.7441
	(0.5,0.6]	20.7836	20.3978	20.5042	21.3015	**21.5946**
SSIM↑	(0.1,0.2]	0.9289	0.9424	**0.9505**	0.9434	0.9489
	(0.2,0.3]	0.8875	0.8916	0.8934	0.9083	**0.9149**
	(0.3,0.4]	0.8437	0.8369	0.8527	0.8735	**0.8809**
	(0.4,0.5]	0.7932	0.7767	0.8032	0.8351	**0.8416**
	(0.5,0.6]	0.6769	0.6664	0.6937	0.7213	**0.7280**
MAE↓	(0.1,0.2]	0.0144	0.0126	0.0118	0.0238	**0.0094**
	(0.2,0.3]	0.0245	0.0233	0.0215	0.0296	**0.0188**
	(0.3,0.4]	0.0345	0.0344	0.0319	0.0423	**0.0266**
	(0.4,0.5]	0.0453	0.0476	0.0438	0.0563	**0.0364**
	(0.5,0.6]	0.0647	0.0664	0.0633	0.0738	**0.0548**
FID↓	(0.1,0.2]	25.0562	25.1855	**14.5851**	18.5755	15.0556
	(0.2,0.3]	53.5397	53.9972	42.0211	39.6055	**28.6686**
	(0.3,0.4]	63.9667	81.1525	53.7013	48.0577	**42.3663**
	(0.4,0.5]	87.0932	103.0111	49.9758	54.4916	**53.8751**
	(0.5,0.6]	106.2307	135.0156	76.4924	80.6035	**73.1309**

Table 2. These are the experimental data of the proposed image inpainting method in this paper with several other image inpainting methods on the CelebA dataset, where the second column represents the percentage of damaged area in the damaged image.

Metrics		**Methods**				
		FFTI [44]	**SWT [45]**	**FMGDN [46]**	**MFMAM [3]**	**Proposed**
PSNR↑	(0.1,0.2]	29.1628	29.7304	30.1550	29.9982	**31.0290**
	(0.2,0.3]	26.1752	26.3189	26.9222	27.4576	**28.4681**
	(0.3,0.4]	24.7306	24.4325	24.5076	25.2564	**25.8856**
	(0.4,0.5]	23.3957	22.8416	22.7389	**24.2805**	23.9599
	(0.5,0.6]	21.0543	20.6740	20.7684	21.5753	**22.1687**
SSIM↑	(0.1,0.2]	0.9325	0.9464	0.9685	0.9699	**0.9797**
	(0.2,0.3]	0.8968	0.9059	0.9107	0.9310	**0.9436**
	(0.3,0.4]	0.8489	0.8476	0.8700	0.8844	**0.9056**
	(0.4,0.5]	0.7966	0.7807	0.8233	0.8503	**0.8705**
	(0.5,0.6]	0.6845	0.6826	0.7099	0.7370	**0.7582**
MAE↓	(0.1,0.2]	0.0117	0.0115	0.0094	0.0203	**0.0072**
	(0.2,0.3]	0.0197	0.0199	0.0188	0.0270	**0.0165**
	(0.3,0.4]	0.0289	0.0290	0.0274	0.0386	**0.0240**
	(0.4,0.5]	0.0407	0.0430	0.0401	0.0540	**0.0343**
	(0.5,0.6]	0.0596	0.0626	0.0603	0.0711	**0.0523**
FID↓	(0.1,0.2]	22.1442	24.5425	**14.9854**	19.6536	15.9213
	(0.2,0.3]	49.1232	45.4356	42.0032	37.1593	**29.1399**
	(0.3,0.4]	66.3441	67.2370	50.1346	48.4633	**41.3119**
	(0.4,0.5]	85.5922	89.0812	74.1722	69.4908	**61.7701**
	(0.5,0.6]	100.4287	115.9153	92.4205	87.0101	**75.6314**

4.3. Qualitative Comparison

In image inpainting tasks, compared to achieving high scores on various metrics, the visual quality and authenticity of the restoration results are more crucial. Given that image inpainting currently cannot fully restore areas with extensive damage, it is essential that the filled content in these missing areas appears as visually authentic and coherent as possible. Therefore, image inpainting often requires qualitative assessment. Figures 5 and 6 present the experimental results of the image inpainting method proposed in this paper, as well as several other advanced methods, on the Paris Street View and CelebA datasets, respectively. Observations from the Paris Street View dataset reveal that our structure-assisted guidance repair method can restore more structural details compared to other methods while ensuring the accuracy of the predicted structures. This is due to our proposed dual U-Net constrained edge prediction network, which can provide more accurate constraints during training to enhance the prediction capability of the edge prediction branch. During the subsequent stage of image texture and semantic content prediction, this structural guidance allows the network to generate images with more accurate lines and fewer structural distortions, even when most of the image is missing. The experimental results also confirm our method's significant advantages in handling structurally dense image inpainting tasks. In the CelebA facial dataset, where structural recovery demands are not as strict but the accuracy of semantic restoration is crucial, the subtleties of facial features, such as facial contours, hair textures, and the positioning of eyes and lips, require the image inpainting algorithm to have a strong capability for semantic extraction and generation. Our proposed multi-scale attention pyramid feature fusion module plays a crucial role in this task, focusing on semantic details to ensure that the restored facial features accurately match the real human appearance. From the experimental results, our method can restore more hair texture and facial details compared to other methods, significantly surpassing others in terms of skin smoothness and the natural appearance of the eyes.

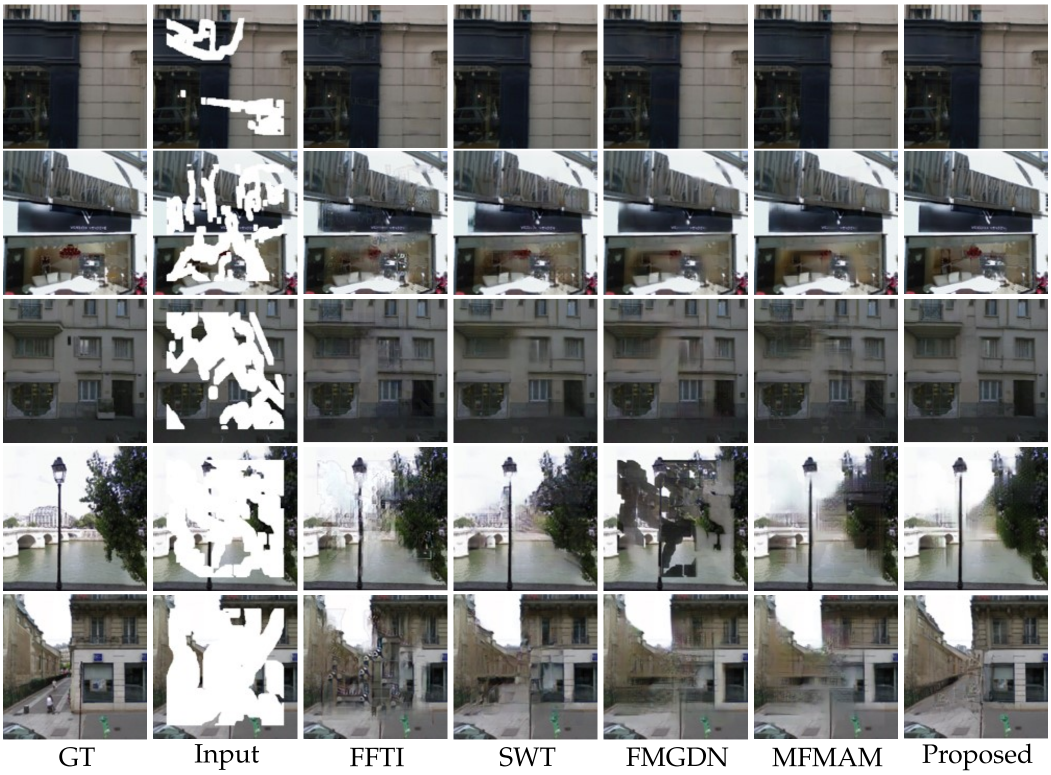

Figure 5. Inpainting results of different image inpainting methods on the Paris Street View dataset. The first column presents the real image, the second column shows the damaged image, and the subsequent columns display the inpainting results of FFTI, SWT, FMGDN, MFMAM, and the proposed method, respectively.

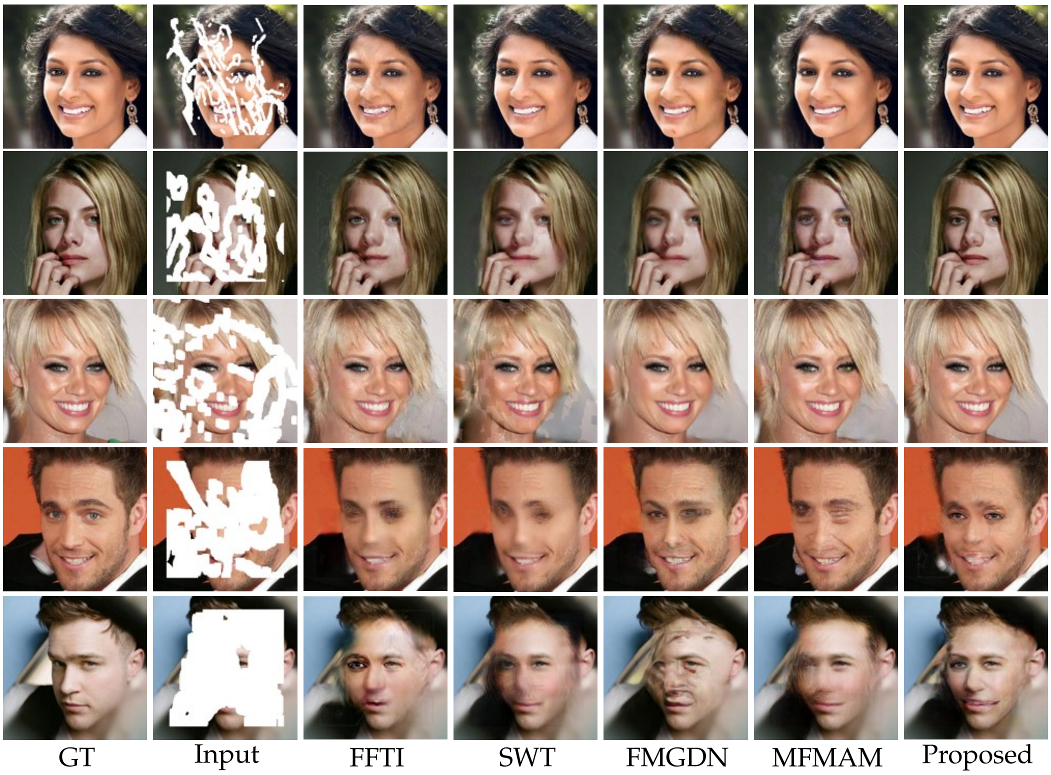

Figure 6. Inpainting results of different image inpainting methods on the CelebA dataset. The first column presents the real image, the second column shows the damaged image, and the subsequent columns display the inpainting results of FFTI, SWT, FMGDN, MFMAM, and the proposed method, respectively.

4.4. Ablation Studies

To substantiate the efficacy of the individual components within the proposed structure-guided multi-scale attention pyramid network, ablation studies were conducted in this section. The baseline was initially established as a standard pyramid inpainting network without incorporating any of the components proposed in this paper. In *Test 1*, the network was modified into a dual-branch structure with edge-assisted repair to improve structural accuracy. *Test 2* further enhanced this configuration by incorporating the multi-scale attention feature fusion module, allowing for better integration of features across different scales. In *Test 3*, part of the conventional convolutions in the network was replaced with separable masked update convolutions to refine the inpainting process.

The overall network was optimized with hyperparameters set to $\lambda_{mul} = 10$, $\lambda_r = 10$, $\lambda_{adv} = 1$, and $\lambda_{fm} = 90$ to better generate detailed textures and structures. These parameters were particularly effective on datasets with complex structures, such as the Paris Street View dataset, where they demonstrated superior performance.

Experimental results, as depicted in Table 3, indicate that the addition of each new module led to improved performance metrics over the previous baseline. This progression underscores the significant contribution of each component to the enhancement of image inpainting outcomes. Notably, the introduction of separable masked update convolutions in the proposed network resulted in a significant reduction in network parameters, with a decrease of approximately 28%. This reduction highlights the critical contribution of the separable masked update convolutions in optimizing large-scale image inpainting networks. The gated mechanism within these convolutions effectively mitigates the impact of erroneous information during the inpainting process, leading to improved restoration accuracy. Moreover, this mechanism not only enhances the overall performance but also substantially reduces computational resource consumption, making it a valuable advancement in the field of image inpainting.

Table 3. This is the experimental data of the proposed image inpainting method in this paper with several other image inpainting methods on the Paris Street View dataset, where the second column represents the percentage of damaged area in the damaged image. The 'M' here stands for 'million', representing the number of parameters that need to be learned in the network.

Metrics		**Methods**			
		Baseline	**Test1**	**Test2**	**Test3**
PSNR↑	(0.1,0.2]	30.3147	31.4310	32.9230	**33.4027**
	(0.2,0.3]	27.3137	27.6520	28.6176	**29.9143**
	(0.3,0.4]	25.2185	25.9490	26.5112	**27.7393**
	(0.4,0.5]	23.6249	23.9725	24.8427	**25.9754**
	(0.5,0.6]	21.0041	21.3820	21.7296	**22.1812**
SSIM↑	(0.1,0.2]	0.9310	0.9473	0.9572	**0.9605**
	(0.2,0.3]	0.8897	0.8977	0.9066	**0.9124**
	(0.3,0.4]	0.8204	0.8544	0.8597	**0.8701**
	(0.4,0.5]	0.7882	0.7933	0.8067	**0.8196**
	(0.5,0.6]	0.6442	0.6631	0.6793	**0.7067**
MAE↓	(0.1,0.2]	0.0169	0.0156	0.0139	**0.0124**
	(0.2,0.3]	0.0288	0.0258	0.0244	**0.0223**
	(0.3,0.4]	0.0396	0.0353	0.0354	**0.0322**
	(0.4,0.5]	0.0463	0.0461	0.0402	**0.0342**
	(0.5,0.6]	0.0710	0.0658	0.0621	**0.0586**
FID↓	(0.1,0.2]	30.9133	27.0090	26.0453	**15.5520**
	(0.2,0.3]	54.8700	49.4695	48.9542	**29.4313**
	(0.3,0.4]	58.6864	52.8904	47.9085	**41.2813**
	(0.4,0.5]	80.6867	74.0636	66.7667	**61.8977**
	(0.5,0.6]	89.2369	83.0700	79.8606	**75.9387**
Parameters↓		**19.3M**	30.9M	33.2M	23.9M

5. Conclusions and Future Work

Current single-view restoration approaches suffer from low image information utilization and subpar repair effects. To enhance image inpainting quality, this paper introduces a structure-guided multi-scale attention pyramid network for image inpainting. This network comprises a structure repair network and a multi-scale attention pyramid semantic repair network. The former achieves constrained structure prediction through a dual-branch U-Net, and the generated structural view serves as auxiliary information for the latter. The latter network exploits the pyramid structure to extract multi-scale features of the image, facilitating feature interaction within the attention feature fusion module. Moreover, the separable gated convolution strategy used during feature extraction minimizes the impact of invalid information in missing regions, ultimately generating accurate and reasonable features for high-quality image inpainting.

In the experimental section, widely-used benchmark datasets such as Paris Street View and CelebA were employed. Quantitative and qualitative comparisons with various methods demonstrate the superiority of the proposed method over existing approaches. Ablation studies were conducted by setting a baseline model and incrementally adding the mechanisms proposed in this paper, comparing each step's results to the new baseline. The analysis of results conclusively demonstrates the effectiveness of the proposed multi-view restoration strategy, separable gated convolution strategy, and multi-scale attention feature fusion strategy.

However, despite these advancements, the proposed method does face several challenges. Maintaining consistent texture and structural details in the inpainted regions remains difficult, particularly when dealing with high-contrast edges or large missing areas. The boundaries between the inpainted area and the surrounding regions can sometimes exhibit artifacts, such as visible seams or color mismatches. Moreover, ensuring that the inpainted region aligns semantically with the surrounding content is a significant challenge, especially in complex scenes like facial images or intricate urban landscapes.

Additionally, while the dual-branch U-Net architecture improves structure prediction, its complex design may lead to increased computational demands, making it less suitable for real-time applications or deployment on devices with limited resources. Training stability and convergence can also be challenging due to the intricate architecture, requiring careful tuning of hyperparameters. Furthermore, the computational expense of multi-scale attention mechanisms, due to processing information at various scales, can result in longer training and inference times.

Future work will focus on addressing these limitations. First, we aim to explore lightweight and more efficient network architectures that retain the performance benefits of the current model while reducing computational complexity. This may involve integrating parallel computation techniques as discussed in related works, such as "Real-time Automated Image Segmentation Technique for Cerebral Aneurysm on Reconfigurable System-On-Chip" [47]. Additionally, we plan to incorporate advanced regularization techniques and preprocessing steps to enhance the model's robustness and reduce artifacts at the boundaries of inpainted regions.

Finally, we intend to optimize the proposed network for deployment on computationally constrained devices, including embedded systems, to broaden its application potential. This will involve exploring techniques to further reduce model size and improve inference speed without sacrificing quality. By addressing these challenges, we hope to make significant strides in the practical applicability and effectiveness of image inpainting frameworks, particularly in scenarios requiring high levels of detail and accuracy.

Author Contributions: Conceptualization, J.G. and S.L.; methodology, J.G.; validation, J.G., S.L. and W.Y.; formal analysis, J.G.; investigation, J.G.; data curation, W.Y.; writing—original draft preparation, J.G.; writing—review and editing, J.G.; visualization, J.G.; supervision, L.N.; project administration, L.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://drive.google.com/file/d/1v6pag_8uL6FOlw8KillKQvKtX-uYbm5u/view?usp=sharing (accessed on 7 September 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

PDE	Partial Differential Equation
E2I	Edge-to-Image
HED	Holistically Nested Edge Detection
CNN	Convolutional Neural Network
GAN	Generative Adversarial Network

References

1. Corneanu, C.; Gadde, R.; Martinez, A.M. LatentPaint: Image Inpainting in Latent Space with Diffusion Models. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2024; pp. 4334–4343.
2. Huang, W.; Deng, Y.; Hui, S.; Wu, Y.; Zhou, S.; Wang, J. Sparse self-attention transformer for image inpainting. *Pattern Recognit.* **2024**, *145*, 109897. [CrossRef]
3. Chen, Y.; Xia, R.; Yang, K.; Zou, K. MFMAM: Image inpainting via multi-scale feature module with attention module. *Comput. Vis. Image Underst.* **2024**, *238*, 103883. [CrossRef]
4. Zhang, X.; Zhai, D.; Li, T.; Zhou, Y.; Lin, Y. Image inpainting based on deep learning: A review. *Inf. Fusion* **2023**, *90*, 74–94. [CrossRef]

5. Zhang, K.; Mo, L.; Chen, W.; Sun, H.; Su, Y. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Adv. Neural Inf. Process. Syst.* **2024**, *36*.
6. Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; Irani, M. Imagic: Text-based real image editing with diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 6007–6017.
7. Yildirim, A.B.; Pehlivan, H.; Bilecen, B.B.; Dundar, A. Diverse inpainting and editing with gan inversion. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 4–6 October 2023; pp. 23120–23130.
8. Gaber, J.A.; Youssef, S.M.; Fathalla, K.M. The Role of Artificial Intelligence and Machine Learning in preserving Cultural Heritage and Art Works via Virtual Restoration. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2023**, *10*, 185–190. [CrossRef]
9. Cai, W.; Xu, X.; Xu, J.; Zhang, H.; Yang, H.; Zhang, K.; He, S. Hierarchical damage correlations for old photo restoration. *Inf. Fusion* **2024**, *107*, 102340. [CrossRef]
10. Chauhan, K.; Patel, S.N.; Kumhar, M.; Bhatia, J.; Tanwar, S.; Davidson, I.E.; Mazibuko, T.F.; Sharma, R. Deep learning-based single-image super-resolution: A comprehensive review. *IEEE Access* **2023**, *11*, 21811–21830. [CrossRef]
11. Chen, Y.; Xia, R.; Yang, K.; Zou, K. MFFN: Image super-resolution via multi-level features fusion network. *Vis. Comput.* **2024**, *40*, 489–504. [CrossRef]
12. Kumar, N.; Meenpal, T. Encoder–decoder-based CNN model for detection of object removal by image inpainting. *J. Electron. Imaging* **2023**, *32*, 042110. [CrossRef]
13. Wei, F.; Funkhouser, T.; Rusinkiewicz, S. Clutter Detection and Removal in 3D Scenes with View-Consistent Inpainting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 4–6 October 2023; pp. 18131–18141.
14. Criminisi, A.; Pérez, P.; Toyama, K. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans. Image Process.* **2004**, *13*, 1200–1212. [CrossRef]
15. Barnes, C.; Shechtman, E.; Finkelstein, A.; Goldman, D.B. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.* **2009**, *28*, 24. [CrossRef]
16. Bertalmio, M.; Sapiro, G.; Caselles, V.; Ballester, C. Image inpainting. In Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, New Orleans, LA, USA, 23–28 July 2000; pp. 417–424.
17. Chan, T.F.; Shen, J. Nontexture inpainting by curvature-driven diffusions. *J. Vis. Commun. Image Represent.* **2001**, *12*, 436–449. [CrossRef]
18. Xu, Z.; Lian, X.; Feng, L. Image inpainting algorithm based on partial differential equation. In Proceedings of the 2008 ISECS International Colloquium on Computing, Communication, Control, and Management, Guangzhou, China, 3–4 August 2008; IEEE: Piscataway, NJ, USA, 2008; Volume 1, pp. 120–124.
19. Li, L.; Dou, Z.Y.; Peng, N.; Chang, K.W. Desco: Learning object recognition with rich language descriptions. *Adv. Neural Inf. Process. Syst.* **2024**, *36*.
20. Jianfeng, L.; Zhongliang, Y.; Yifan, L.; Guanghui, S. GTAN: Graph-based tracklet association network for multi-object tracking. *Neural Comput. Appl.* **2024**, *36*, 3889–3902. [CrossRef]
21. Wang, Q.; Sun, H.; Feng, Y.; Dong, Z.; Bai, C. MGCNet: Multi-granularity cataract classification using denoising diffusion probabilistic model. *Displays* **2024**, *83*, 102716. [CrossRef]
22. Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context Encoders: Feature Learning by Inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
23. Min, W.; Fan, M.; Guo, X.; Han, Q. A new approach to track multiple vehicles with the combination of robust detection and two classifiers. *IEEE Trans. Intell. Transp. Syst.* **2017**, *19*, 174–186. [CrossRef]
24. Wang, Q.; Min, W.; He, D.; Zou, S.; Huang, T.; Zhang, Y.; Liu, R. Discriminative fine-grained network for vehicle re-identification using two-stage re-ranking. *Sci. China Inf. Sci.* **2020**, *63*, 1–12. [CrossRef]
25. Zhao, H.; Min, W.; Xu, J.; Han, Q.; Li, W.; Wang, Q.; Yang, Z.; Zhou, L. SPACE: Finding key-speaker in complex multi-person scenes. *IEEE Trans. Emerg. Top. Comput.* **2021**, *10*, 1645–1656. [CrossRef]
26. Liu, G.; Reda, F.A.; Shih, K.J.; Wang, T.C.; Tao, A.; Catanzaro, B. Image inpainting for irregular holes using partial convolutions. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 85–100.
27. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Free-form image inpainting with gated convolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4471–4480.
28. Liu, Q.; Li, S.; Xiao, J.; Zhang, M. Multi-filters guided low-rank tensor coding for image inpainting. *Signal Process. Image Commun.* **2019**, *73*, 70–83.
29. Wei, Z.; Min, W.; Wang, Q.; Liu, Q.; Zhao, H. ECNFP: Edge-constrained network using a feature pyramid for image inpainting. *Expert Syst. Appl.* **2022**, *207*, 118070. [CrossRef]
30. Xiong, W.; Yu, J.; Lin, Z.; Yang, J.; Lu, X.; Barnes, C.; Luo, J. Foreground-aware image inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5840–5848.
31. Hedjazi, M.A.; Genc, Y. Efficient texture-aware multi-GAN for image inpainting. *Knowl.-Based Syst.* **2021**, *217*, 106789. [CrossRef]
32. Chen, Y.; Xia, R.; Yang, K.; Zou, K. DNNAM: Image Inpainting Algorithm via Deep Neural Networks and Attention Mechanism. *Appl. Soft Comput.* **2024**, *154*, 111392. [CrossRef]
33. Bai, J.; Fan, Y.; Zhao, Z.; Zheng, L. Image Inpainting Technique Incorporating Edge Prior and Attention Mechanism. *Comput. Mater. Contin.* **2024**, *78*. [CrossRef]

34. Xiang, H.; Min, W.; Wei, Z.; Zhu, M.; Liu, M.; Deng, Z. Image inpainting network based on multi-level attention mechanism. *IET Image Process.* **2024**, *18*, 428–438. [CrossRef]
35. Iizuka, S.; Simo-Serra, E.; Ishikawa, H. Globally and Locally Consistent Image Completion. *ACM Trans. Graph.* **2017**, *36*, 1–14. [CrossRef]
36. Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.Z.; Ebrahimi, M. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv* **2019**, arXiv:1901.00212.
37. Xu, S.; Liu, D.; Xiong, Z. E2I: Generative inpainting from edge to image. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 1308–1322. [CrossRef]
38. Ren, Y.; Yu, X.; Zhang, R.; Li, T.H.; Liu, S.; Li, G. Structureflow: Image inpainting via structure-aware appearance flow. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 181–190.
39. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Generative Image Inpainting with Contextual Attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
40. Zheng, C.; Cham, T.J.; Cai, J. Pluralistic image completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1438–1447.
41. Zeng, Y.; Fu, J.; Chao, H.; Guo, B. Learning pyramid-context encoder network for high-quality image inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1486–1494.
42. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
43. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
44. Chen, Y.; Xia, R.; Zou, K.; Yang, K. FFTI: Image inpainting algorithm via features fusion and two-steps inpainting. *J. Vis. Commun. Image Represent.* **2023**, *91*, 103776. [CrossRef]
45. Chen, B.W.; Liu, T.J.; Liu, K.H. Lightweight Image Inpainting by Stripe Window Transformer with Joint Attention to CNN. In Proceedings of the 2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP), Rome, Italy, 17–20 September 2023; pp. 1–6.
46. Zhang, X.; Hamann, B.; Wang, D.; Wang, H.; Wang, Y.; Yin, Y.; Gao, H. FMGDN: Flexible Multi-Grained Dilation Network Empowered Multimedia Image Inpainting for Electronic Consumer. *IEEE Trans. Consum. Electron.* **2024**, *70*, 4816–4827. [CrossRef]
47. Zhai, X.; Eslami, M.; Hussein, E.S.; Filali, M.S.; Shalaby, S.T.; Amira, A.; Bensaali, F.; Dakua, S.; Abinahed, J.; Al-Ansari, A.; et al. Real-time automated image segmentation technique for cerebral aneurysm on reconfigurable system-on-chip. *J. Comput. Sci.* **2018**, *27*, 35–45. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.