蘭州理工大學

科研成果汇总

学　　号： 201081104001

研 究 生： 史长宏

导　　师： 刘微容　教授

研究方向： 图像处理与模式识别

论文题目： 多模态注意力机制下的图像修复与超分辨率方法研究

学　　科： 模式识别与智能系统

学　　院： 自动化与电气工程学院

入学时间： 2020 年 9 月

目　录

图书馆 文献检索报告

兰州理工大学图书馆 报告编号：R2025-0786

机构：兰州理工大学 电气工程与信息工程学院

姓名：史长宏 [201081104001]

著者要求对其在国内外学术出版物所发表的科技论著被以下数据库收录情况进行查证。

检索范围：

- 科学引文索引（Science Citation Index Expanded）：1900年-2025年
- 工程索引（Engineering Index）：1884年-2026年

检索结果：

检索类型	数据库	年份范围	总篇数	第一作者篇数
SCI-E 收录	SCI-EXPANDED	2025	1	1
EI 收录	EI-Compendex	2025 - 2026	2	2

委托人声明：

本人委托兰州理工大学图书馆查询论著被指定检索工具收录情况，经核对检索结果，附件中所列文献均为本人论著，特此声明。

作　　者（签字）：史长宏

完成人（签字）：刘　玮

完　成　日　期　：2025年9月12日

完成单位（盖章）：兰州理工大学图书馆信息咨询与学科服务部

（本检索报告仅限校内使用）

图书馆

文献检索报告
SCI-E 收录

兰州理工大学图书馆

报告编号：R2025-0786 SCI-E 收录

数据库：科学引文索引 (Science Citation Index Expanded) 时间范围：2025年	作者姓名：史长宏 作者单位：兰州理工大学 电气工程与信息工程学院	检索人员：刘 玮 检索日期：2025年9月12日

检索结果：被 SCI-E 收录文献 1 篇

#	作者	地址	标题	来源出版物	文献类型	入藏号
1	**Shi, CH**; Liu, WR; Meng, JH; Jia, XF; Liu, J	[Shi, Changhong; Liu, Weirong; Meng, Jiahao; Jia, Xiongfei; Liu, Jie] Lanzhou Univ Technol, Coll Elect & Informat Engn, Langongping Rd, Qilihe Dist, Lanzhou 730050, Gansu, Peoples R China	Self-prior guided generative adversarial network for image inpainting	*VISUAL COMPUTER* 2025, 41 (4): 2939-2951.	J Article	WOS:001283873100001
					合计	1

图书馆

兰州理工大学图书馆

文献检索报告
EI 收录

报告编号：R2025-0786 EI 收录

数据库：工程索引 (Engineering Index) 时间范围：2025年至2026年	作者姓名：史长宏 作者单位：兰州理工大学 电气工程与信息工程学院	检索人员：刘 玮 检索日期：2025年9月12日

检索结果：被 EI 收录文献 2 篇

#	作者	地址	标题	来源出版物	文献类型	入藏号
1	**Shi, Changhong**; Liu, Weirong; Li, Zhijun; Yi, Jiajing; Liu, Jie	College of Electric and Informational Engineering, Lanzhou University of Technology, Gansu, LanZhou	Cultural Relic Image Inpainting via Multi-column Condition Decoding Transformer	***Sensing and Imaging*** 2025, 26 (1): 98.	Journal article (JA)	20252818745402
2	**Shi, Changhong**; Liu, Weirong; Meng, Jiahao; Li, Zhijun; Liu, Jie	College of Electric and Informational Engineering, Lanzhou University of Technology, Gansu, Lanzhou	Global Cross Attention Transformer for Image Super-Resolution	***Lecture Notes in Computer Science*** 2026, 15951 LNCS: 161-171.	Conference article (CA)	20253519086483
					合计	2

The Visual Computer (2025) 41:2939–2951
https://doi.org/10.1007/s00371-024-03578-x

RESEARCH

Self-prior guided generative adversarial network for image inpainting

Changhong Shi[1] · Weirong Liu[1] · Jiahao Meng[1] · Xiongfei Jia[1] · Jie Liu[1]

Accepted: 11 July 2024 / Published online: 5 August 2024

Abstract
Great progress has been made in image inpainting tasks with the emergence of convolutional neural networks, because of their superior translation invariance and powerful texture modeling capacity. However, current solutions generally do not perform well in reconstructing high-quality results. To address this issues, a self-prior guided generative adversarial network (SG-GAN) model is proposed. SG-GAN integrates the learning paradigms of cross-attention and convolution to the generator. It is able to learn the cross-mapping between input and target dataset effectively. Then, a high receptive field subnet is constructed to increase the receptive field. Finally, a high receptive field feature-matching loss is proposed to further ensure the structure sharpness of generated images. Experiments on datasets including natural scene images (Places2), facial images (CelebA-HQ), structured wall images (Façade), and Dunhuang Mural images show that the proposed method can generate higher quality results with more details than state-of-the-art.

Keywords Image inpainting · Generative adversarial networks · Cross attention · High receptive field · Feature-matching loss

1 Introduction

Image inpainting aims to hallucinate missing regions with semantically reasonable and visually realistic content by using the known regions (background regions) [1]. It is one of the key researches in low-level visual tasks, and it is needed in practical applications such as image dehazing [2], image recognition [3, 4], face editing [5–7], precious historical data restoration [8, 9], and image decomposition [10]. Earlier methods [11, 12] fill missing regions by disseminating local information about neighboring visible regions or searching for matching patches. These methods usually generate unreasonable results without considering the semantic information.

Recently, progress has been accelerated through the use of deep neural networks and adversarial learning thoughts in a data-driven manner [13–17]. For example, CE [18] is the earliest image inpainting method that fills regular invisible regions by adopting an encoder-decoder structure. However, images generated by context encoder often contain obvious visual artifacts. To address this issue, CA [1] employs an embeddable attention module to transfer features of undamaged regions to missing regions. Unfortunately, the above improvement is insufficient to generate reasonable semantics. The attention mechanism requires roughly completed features of undamaged regions. However, this becomes difficult to satisfy as the missing area becomes larger. Therefore, complicated models with intermediate predictions have been designed to address the above issue. For example, two-stage or multi-stage progressive networks [19–23], in which the propagation of long-range dependences is achieved in a deep feature extraction process by convolution. Nevertheless, it exists a challenge to learn adaptive matching weights to realize long-range dependences in terms of spatial feature extraction and propagation.

Jiahao Meng, Xiongfei Jia and Jie Liu have contributed equally to this work.

✉ Weirong Liu
liuwr@lut.edu.cn

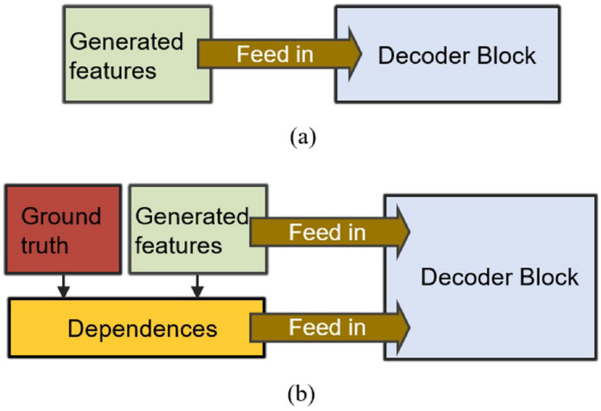
Changhong Shi
changhong_shi@126.com

Jiahao Meng
mjhforwork@163.com

Xiongfei Jia
jiaxiongfei_lut@163.com

Jie Liu
ljdaisy@163.com

[1] College of Electric and Informational Engineering, Lanzhou University of Technology, Langongping Road, Qilihe District, LanZhou 730050, Gansu, China



(a)

(b)

Fig. 1 Illustration of input forms participating in decoder

Thanks to the development of self-attention [24, 25], transformer-based architectures [26–28] have emerged to solve the above issues. These approaches apply the transformer to the encoder or small scale layers of a generator. However, only the hidden layer feature is considered as input in the forward propagation process of decoder network parameters, ignoring the explicit dependency relationship between hidden layer features and ground truth images, as shown in Fig. 1. Therefore, the learning ability of decoders is limited, which affects the quality of generated images.

An image inpainting model called SG-GAN has been developed to resolve the above challenges. Firstly, the cross-attention mechanism is added to the decoder, which enables the generator to learn an effective mapping by utilizing cross-dependency between ground truth images and hidden layer features. Specifically, ground truth images are introduced into a decoder inspired by a transformer in which ground truth images are leveraged to guide the training process, as shown in Fig. 1. Therefore, the traditional generator network has been improved to create a unique fusion learning paradigm of convolution and cross-attention mechanism. Convolution layer and cross-attention layer (CrA) form decoder block called cross-attention and convolution (CAC), as shown on the left of Fig. 2.

To further improve the quality of image inpainting results, it is necessary to address the structural discontinuity caused by the absence of a receptive field amplification mechanism. To this end, an HRF subnet is constructed, and corresponding HRFM loss is proposed to determine activation features of the dilated convolutional layer in the HRF subnet. The proposed HRFM loss provides a new criterion for the structural consistency of the completion model. The level of structural consistency is enhanced in the final completed images by leveraging HRFM loss.

Experiments conducted on natural scene images (Places2) [29], structured wall images (Façade) [30], and Dunhuang Mural images [31] validate that the proposed method can generate plausible results. The proposed method outperforms existing state-of-the-art both quantitatively and qualitatively. In a nutshell, the main contributions are summarized as follows:

- A CAC block composed of CrA and convolution layer is proposed for the decoder construction, which enhances the modeling capacity of generator by leveraging the cross-dependencies of CrA and the powerful texture perception ability of convolution. The proposed CAC block allows the generator to adopt a fusion learning paradigm.
- An HRF subnet is constructed and corresponding HRFM loss is proposed. The HRFM loss distinguishes activation characteristics of dilated convolution layers, providing a criterion for ensuring structural consistency and clarity of generated results.
- Finally, various experiments demonstrate that the proposed SG-GAN architecture outperforms the existing state-of-the-art image inpainting methods.

2 Related work

2.1 Image inpainting by CNN-based models

Since Pathak et al. [18] introduced adversarial learning into the image inpainting task, a series of improved methods [1, 19, 32–35] for generators and discriminators have been explored to reconstruct meaningful content. For example, MC [36] is a generative multi-column model, which synthesizes different image components exploiting different scalers of the filter in a parallel way. PEN-Net [37] learns region affinity utilizing an attention mechanism and fills missing regions from small scaler to large scaler in a pyramid structure. GConv [38] introduces a two-branch refinement network to learn a dynamic feature selection mechanism and presents a practical patch-based GAN discriminator, which is followed by [39–43]. EdgeConnect [19] designs an texture generator followed by an edge generator as structural guidance for completing missing structures and textures in two stages, which is followed by StructureFlow [20]. GL [44] employs global and local discriminators to improve image quality and get better consistency around the boundary of missing regions. Most of these models are built upon a CNN-based encoder-decoder architecture, in which there exists the issue of a single learning paradigm and can hardly generate sharp structures due to a lack of receptive field amplification mechanism. In contrast, the proposed model builds a generator with a fusion learning paradigm and designs an HRF subnet to ensure the generation of natural texture and clear structure.

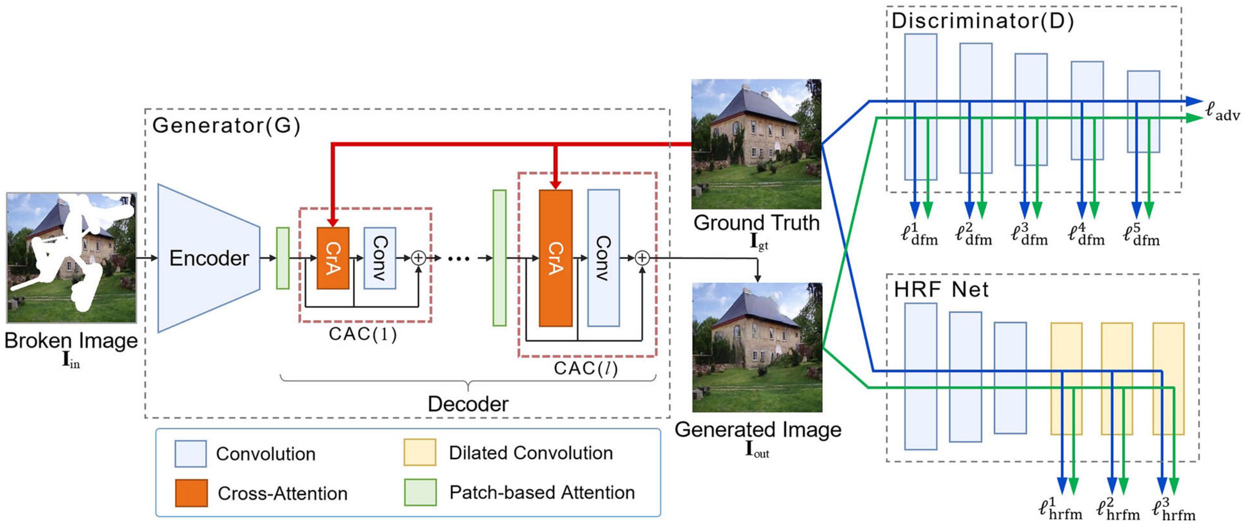


Fig. 2 Overview of the proposed SG-GAN framework. The framework contains three parts: the generator network G, the discriminator network D, and the HRF subnetwork. Inputs of the framework are broken image $\mathbf{I}_{\text{in}}$, and outputs of G are generated image $\mathbf{I}_{\text{out}}$. The ground truth images $\mathbf{I}_{\text{gt}}$ participate in the training of the G network. The ground truth images $\mathbf{I}_{\text{gt}}$ and $\mathbf{I}_{\text{out}}$ are fed into D and HRF network. The discrimination results of the discriminator network are fed back to the generator network through generation adversarial loss ℓ_{adv}. ℓ^i_{dfm} and ℓ^i_{hrfm} are feature-matching losses of the D network and HRF network respectively

2.2 Self-attention

Self-attention methods [45] are divided into two categories: non-learning self-attention methods and learnable self-attention methods. Non-learning self-attention methods [46, 47] embed self-attention modules into convolution networks. However, the embedded self-attention module calculates long-range dependencies in the hidden layer feature map with CNN network in these methods, but struggles to adaptively learn long-range dependencies. Transformer, a learnable self-attention method for machine translation, was proposed by Vaswani et al. [24] to address the above limitations. The transformer adopts an encoder-decoder architecture consisting of multiple learnable self-attention layers. Recent works have applied transformers to visual tasks [48, 49], such as image classification [27, 50, 51], image generation [22], and image completion [26, 28]. Encoders of such methods are mostly replaced by encoders of transformers crudely. However, cross-attention mechanisms are one of the keys to the performance of transformer, and the cross-attention mechanism in decoders is not exploited by the above methods. In contrast, a cross-attention mechanism is introduced into the decoder in this manuscript, it exploits cross-dependencies between ground truth images and hidden layer features to learn efficient mappings between inputs and outputs.

3 Self-prior guided GAN

The proposed SG-GAN framework consists of three parts: the generator network G, the discriminator network D, and the HRF subnetwork. Figure 2 for an overview. The proposed architecture follows an adversarial model, i.e., discriminator D is trained to push the generator to reach its goal by distinguishing the authenticity of generated images from G.

3.1 Generator

G consists of an encoder and decoder (Fig. 2). The encoder compresses and encodes damaged images into compact latent features through vanilla convolution and subsampling operations. The decoder up sample latent features back to a complete RGB image. The encoder is composed of a simple convolution layer, and the decoder is composed of a fusion of the convolution layer and CrA. Inspired by the transformer, CrA is introduced to form a hybrid decoding block CAC, so that the SG-GAN framework is equipped with a fusion learning paradigm. Let $\mathbf{I}_{\text{gt}}$ be ground truth images and the input of the generator is $\mathbf{I}_{\text{in}} = \mathbf{I}_{\text{gt}} \odot (1 - \mathbf{M})$. Here, $\odot$ denotes the Hadamard product. Mask $\mathbf{M}$ is a binary region (with a value of 0 for known pixels and 1 for missing regions).

3.1.1 CAC block

In this section, the details of the CAC block (Fig. 3) are illustrated. The first block is used as an example to introduce the calculation process. Let $\mathbf{I}$ be latent hidden features generated

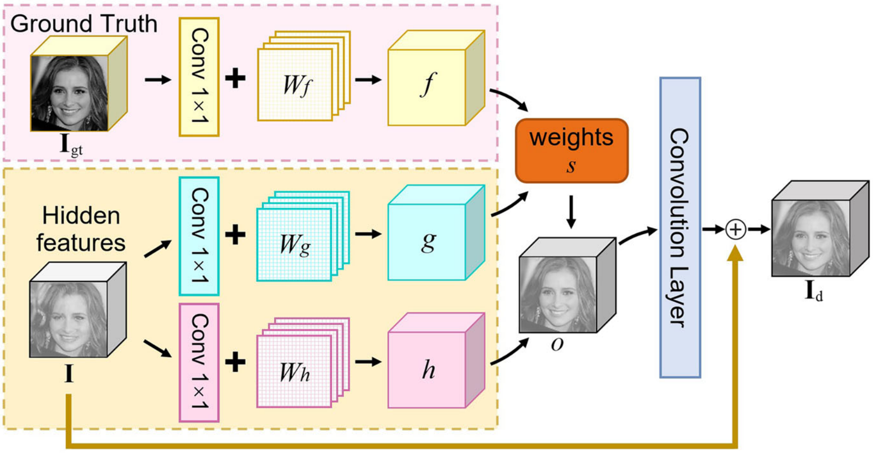


Fig. 3 Flowchart of the CAC block

by the encoder, $\mathbf{I}_d$ be the output of the CAC block. The input of CrA concludes in three parts. One comes from the ground truth and the other two come from the previous hidden layer. These three inputs are performed 1×1 convolution $\mathrm{Conv1}(\cdot)$ and linear operations to get three feature spaces by formulas (1):

$$\begin{aligned} f &= \boldsymbol{W}_f \cdot \mathrm{Conv1}(\mathbf{I}_{\mathrm{gt}}) \\ g &= \boldsymbol{W}_g \cdot \mathrm{Conv1}(\mathbf{I}) \\ h &= \boldsymbol{W}_h \cdot \mathrm{Conv1}(\mathbf{I}) \end{aligned} \tag{1}$$

Cross-attention weights indicate the extent to which the model attends to each pixel of ground truth when synthesizing each pixel of hidden features, and cross-attention weights can be calculated as: $s = \mathrm{soft\,max}(fg^{\mathrm{T}})$. The output of CrA is $o = \boldsymbol{W}_o \cdot s \cdot h$, where $\boldsymbol{W}_f$, $\boldsymbol{W}_g$, $\boldsymbol{W}_h$, and $\boldsymbol{W}_o$ are learnable weight matrices, which are implemented as convolutions. The convolution operation of the decoder is denoted as $\mathrm{Dec}(\cdot)$. The output of the CAC block is calculated by formula (2).

$$\mathbf{I}_d = \mathrm{Concat}(\mathbf{I}, \mathrm{Dec}(o)) \tag{2}$$

The proposed CAC block includes convolution and cross-attention learning strategies. In particular, the CrA mechanism can not only learn the dependence between input images, but also learn the dependence between ground truth images and input images, due to the introduction of the ground truth. The learned network parameters containing the above dependencies are stored in matrices $\boldsymbol{W}_f$, $\boldsymbol{W}_g$, and $\boldsymbol{W}_h$.

3.2 Discriminator

In the SG-GAN model, a spectral normalized Markov discriminator (SN-PatchGAN) [38] is leveraged. It is shown on the upper right of Fig. 2. D includes 5 convolution layers with a kernel size of 5 and stride of 2. The fully connected layer is used at the last layer.

3.3 HRF subnet

The HRF subnet is constructed for calculating HRFM loss to promote sharp structure generation. It includes 3 convolution layers with a kernel size of 5 and stride of 2, tailed by dilated convolution layers to increase the receptive field, focusing on the understanding of global structure. Dilated convolution layers are added after the third convolution layer, with a kernel size of 3 and a dilated rate of 2, 4, and 8, respectively.

3.4 Loss function

The proposed framework is trained with a joint loss, containing adversarial loss, HRFM loss, reconstruction loss, and style loss to synthesize visually realistic and semantically reasonable results. A simple additive form is used for the loss function:

$$\mathcal{L} = \mathcal{L}_{\mathrm{adv}} + \lambda_r \mathcal{L}_{\mathrm{rec}} + \lambda_s \mathcal{L}_{\mathrm{style}} + \lambda_f (\mathcal{L}_{\mathrm{hrfm}} + \mathcal{L}_{\mathrm{dfm}}) \tag{3}$$

where λ_r, λ_s and λ_f are regularity factors to balance contributions of losses.

HRFM Loss The HRFM loss $\mathcal{L}_{\mathrm{hrfm}}$ is proposed to constrain sharp structures. It is obtained by applying feature-matching loss to the activation mapping of dilated convolution layers in the HRF subnet. The HRFM loss stabilizes the training process by forcing the generator to generate structural representations similar to ground truth. The HRFM loss is defined as:

$$\mathcal{L}_{\mathrm{hrfm}} = \sum_{i=1}^{L} \frac{1}{N_i} \left\| \mathrm{HRF}^{(i)}(\mathbf{I}_{\mathrm{gt}}) - \mathrm{HRF}^{(i)}(\mathbf{I}_{\mathrm{out}}) \right\| \tag{4}$$

where $\mathbf{I}_{\text{gt}}$ are ground truth images, and $\mathbf{I}_{\text{out}}$ are outputs of the generator. L indicates the total number of dilated convolution layers in the HRF subnet, N_i is the number of elements in the i'th activation layer, and $\text{HRF}^{(i)}$ is the activation in the i'th layer of the HRF subnet. Similarly, $\mathcal{L}_{\text{dfm}}$ is defined as the feature-matching loss of the discriminator D, which is obtained by applying feature-matching loss to the activation mapping of D.

Adversarial Loss Discriminator is trained to push generator to reach its goal by distinguishing the synthesized images from the ground truth images, and the adversarial loss plays a key role in the game process. The adversarial loss can be expressed as:

$$\mathcal{L}_{\text{adv}} = E_{\mathbf{I}_{\text{gt}} \sim P_{\text{data}}} \log D(\mathbf{I}_{\text{gt}}) + E_{\mathbf{I}_{\text{out}} \sim P_G} \log(1 - D(\mathbf{I}_{\text{out}})) \tag{5}$$

where P_{data} and P_G denote the real data distribution and generated data distribution respectively. $D(\mathbf{I}_{\text{gt}})$ represents the probability that $\mathbf{I}_{\text{gt}}$ came from P_{data}. $D(\mathbf{I}_{\text{out}})$ represents the probability that $\mathbf{I}_{\text{out}}$ came from P_G.

Reconstruction Loss The reconstruction loss includes pixel-level L1 loss and VGG loss. L1 Loss ensures reconstruction accuracy by calculating the similarity of pixels at all scales, and VGG loss constrains information by extracting deep features:

$$\mathcal{L}_{\text{rec}} = \left\| (\mathbf{I}_{\text{gt}} - \mathbf{I}_{\text{out}}) \right\|_1 + \left\| VGG_d(\mathbf{I}_{\text{gt}}) - VGG_d(\mathbf{I}_{\text{out}}) \right\|_1 \tag{6}$$

where $VGG_d(\cdot)$ is a pre-trained CNN network, and d is the feature map of a specific VGG layer.

Style Loss The style loss is introduced to ensure style consistency. An auto-correlation (Gram matrix) on each specific VGG feature map is performed before applying L1 loss. The style loss can be defined as:

$$\mathcal{L}_{\text{style}} = \sum_n \left\| (\psi_n(\mathbf{I}_{\text{out}}))^T \psi_n(\mathbf{I}_{\text{out}}) - (\psi_n(\mathbf{I}_{\text{gt}}))^T \psi_n(\mathbf{I}_{\text{gt}}) \right\|_1 \tag{7}$$

where, $\psi_n(\cdot)$ is the activation map of the n'th selected layer. The training schedule of the SG-GAN network is summarized in Algorithm 1.

4 Experiments

The proposed methods are evaluated by comparing with state-of-the-art in terms of objective quantification and subjective perception. In addition, the ablation study is conducted to evaluate the effectiveness of the proposed SG-GAN architecture.

Algorithm 1 The training process of SG-GAN

```
Require: X: dataset, b: batch size, lr: learning rate
Ensure: Images generated by G
Initialize G and D
while G and D does not converge do
    $\mathbf{I}_{\text{gt}} \sim \text{sample}\{\mathbf{X}\}_b$: randomly choice a batch
    Randomly generate irregular mask $\mathbf{M}$
    $\mathbf{I}_{\text{in}} = \mathbf{I}_{\text{gt}} \odot (1 - \mathbf{M})$: obtain $\mathbf{I}_{\text{in}}$
    Get the latent features $\mathbf{I}$ by encoder
    for $i$ = 1 to $l$ do
        Get patch-based attention feature
        Compute $I_d^l$ with Eq.1 and Eq.2
        Get output: $\mathbf{I}_{\text{out}} = \text{ToRGB}(\mathbf{I}_{\text{d}})$
    end for
    Compute the loss $\mathcal{L}$ with Eq.3
    Update discriminator $D$ with $\mathbf{I}_{\text{gt}}$, $\mathbf{I}_{\text{out}}$, and $\mathcal{L}$
    Update generator $G$ with the loss $\mathcal{L}$
    Update HRFM subnet with the loss $\mathcal{L}$
end while
```

Table 1 Training and testing splits of four datasets

Dataset	Train set	Test set	Total
Places2	327,739	761	328,500
CelebA-HQ	29,700	300	30,000
Façade	556	50	606
Dunhuang Mural	1564	50	1614

4.1 Datasets and compared methods

The proposed SG-GAN model is evaluated with irregular masks on various datasets, including a collection of natural scene images with distinct categories Places2 [29], high-quality face attributes dataset CelebA-HQ, building facade images in various styles from different cities around the world Façade [30], and Dunhuang Mural images [31]. Places2 and CelebA-HQ datasets are used for quantitative comparison and qualitative comparison. Façade and Dunhuang Mural datasets are used for ablation study and application respectively. The number of training data sets and test sets is divided as shown in Table 1. The proposed framework SG-GAN is compared with the following baselines: PEN-Net [37], EC [19], HiFill [33], and GDS [40] using their released codes and models. The link for three test datasets including Places2, CelebA-HQ, and Façade is https://drive.google.com/drive/folders/17xAotx4mlJTFibZXwg5u5zbesK0cujGm?usp=drive_link.

4.2 Implementation details

All models are trained on NVIDIA TITAN GPU (12GB) with batch size of 5. The learning rate is set to be 1×10^{-4}. All methods are tested with the same image size 256×256. The Adam optimizer is used to optimize the model, and momentums are set to $\beta_1 = 0.5$ and $\beta_2 = 0.9$, respectively. The balance

Table 2 The quantitative comparison on Places2 dataset. The best result of each column is boldfaced. ↑ and ↓ represent larger and smaller as better, respectively

Metrics	PSNR$^{\uparrow}$			SSIM$^{\uparrow}$			FID$^{\downarrow}$		
Mask ratio	(0.01, 0.2]	(0.2, 0.4]	(0.4, 0.6]	(0.01, 0.2]	(0.2, 0.4]	(0.4, 0.6]	(0.01, 0.2]	(0.2, 0.4]	(0.4, 0.6]
PEN-Net [37]	26.10	23.44	21.53	0.915	0.868	0.820	32.50	51.70	68.77
EC [19]	28.44	25.97	24.25	0.929	0.857	0.854	15.10	23.34	30.02
HiFill [33]	26.11	23.78	21.97	0.901	0.855	0.808	30.70	47.29	66.96
GDS [40]	26.68	24.33	22.62	0.918	0.876	0.833	24.38	39.39	53.46
SGGAN	**29.10**	**26.08**	**24.32**	**0.937**	**0.897**	**0.860**	**14.97**	**23.21**	**30.01**

Table 3 The quantitative comparison on CelebA-HQ dataset. The best result of each column is boldfaced. ↑ and ↓ represent larger and smaller as better, respectively

Metrics	PSNR$^{\uparrow}$			SSIM$^{\uparrow}$			FID$^{\downarrow}$		
Mask ratio	(0.01, 0.2]	(0.2, 0.4]	(0.4, 0.6]	(0.01, 0.2]	(0.2, 0.4]	(0.4, 0.6]	(0.01, 0.2]	(0.2, 0.4]	(0.4, 0.6]
MC [36]	29.77	26.92	25.01	0.949	0.921	0.895	18.17	32.50	44.06
PEN-Net [37]	27.45	24.71	22.96	0.939	0.906	0.876	33.63	57.93	81.63
EC [19]	31.17	28.61	26.89	0.956	0.934	0.913	**12.40**	19.84	**24.11**
GDS [40]	28.76	26.07	24.34	0.946	0.920	0.894	18.19	31.52	41.55
SGGAN	**31.24**	**28.63**	**26.98**	**0.957**	**0.936**	**0.917**	13.14	**19.77**	26.20

parameter of the loss function is determined as $\lambda_r = 1.2$ and $\lambda_s = \lambda_f = 0.01$ after multiple parameter adjustments. Meanwhile, the kernel sizes, strides, and activation functions are empirically set. The proposed model does not perform specific post-processing steps. The generator is divided into the training phase and the testing phase due to the introduction of CrA. In the training stage, the ground truth and hidden layers participate in the learning of the self-attention layer. In testing phase, the trained generator is used for testing, and inputs of the self-attention layer come only from hidden layer features without ground truth. The irregular masks are adopted for both training and testing. The testing masks can be categorized into 3 categories through hole-to-image area ratios: (0.01,0.2], (0.2,0.4], (0.4,0.6]. The measurement metrics contain Peak Signal-to-Noise Ratio (PSNR), Structural SIMilarity (SSIM), and Frèchet Inception Distance (FID). PSNR measures the L2 distance, SSIM measures structure similarity, and FID is proposed to mimic human perception of similarity.

4.3 Quantitative comparison

Quantitative comparison experiments are conducted to evaluate that the proposed SG-GAN achieves superior average values than most of the state-of-the-art methods on PSNR, SSIM, and FID measurements. Tables 2 and 3 show quantitative evaluation results on Places2 and CelebA-HQ datasets with mask ratios (0.01, 0.2], (0.2, 0.4], and (0.4, 0.6]. Comparison experiments are all conducted with irregular masks.

Table 2 shows that the proposed SG-GAN method reconstructs excellent results and achieves superior scores for all metrics. Especially, SG-GAN achieves an average improvement of 12% in PSNR compared with PEN-Net [37], and 5% in SSIM scores compared with HiFill [33], respectively. In addition, SG-GAN reduces the average FID value by 42% compared to GDS [40] with a average value of 39.08. For the CelebA-HQ dataset, SG-GAN produces excellent results for all measurement metrics as shown in Table 3. In particular, SG-GAN achieves the highest average PSNR and SSIM scores among all of the compared baselines.

4.4 Qualitative comparison

Qualitative comparison experiments are conducted to evaluate whether the proposed SG-GAN is superior to the baseline methods in structure and detail preservation. The qualitative comparison results are obtained by using different baselines on Places2 and CelebA-HQ datasets. Figure 4 shows that results generated by SG-GAN are semantically more reasonable compared to PEN-Net [37], avoiding obvious artifacts that PEN-Net [37] often generates (such as lights in the fourth row). Compared with results generated by HiFill [33], SG-GAN generates more natural and smoother results, while HiFill [33] seems to restore blur and distortion (such as the marked regions in the third row). Even when mask regions are complex, SG-GAN is still able to reconstruct the linear ladder area. EC [19] and GDS [40] have superior performance when the missing region have simple content, they

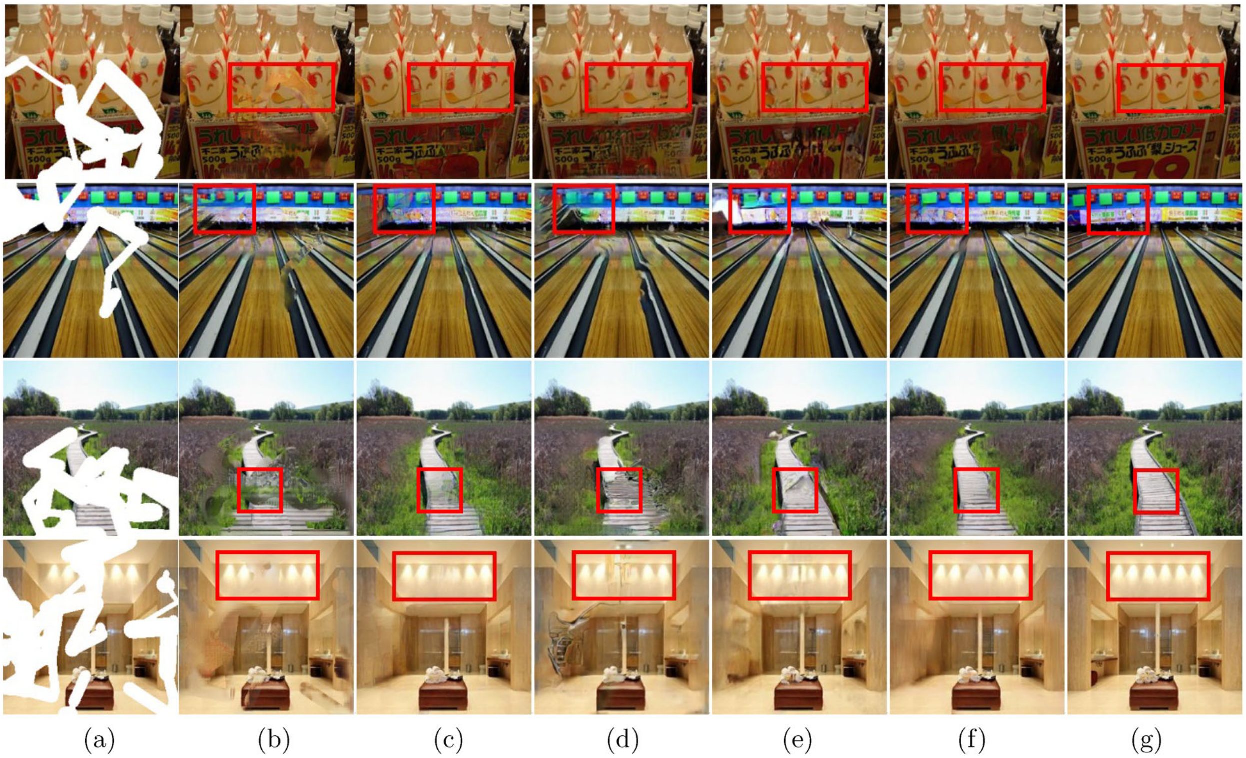


Fig. 4 Qualitative comparisons on Places2 dataset (zoom in for a better view). In each row, images from left to right: **a** Masked inputs, **b** PEN-Net [37], **c** EC [19], **d** HiFill [33], **e** GDS [40], **f** SG-GAN, **g** Ground truth

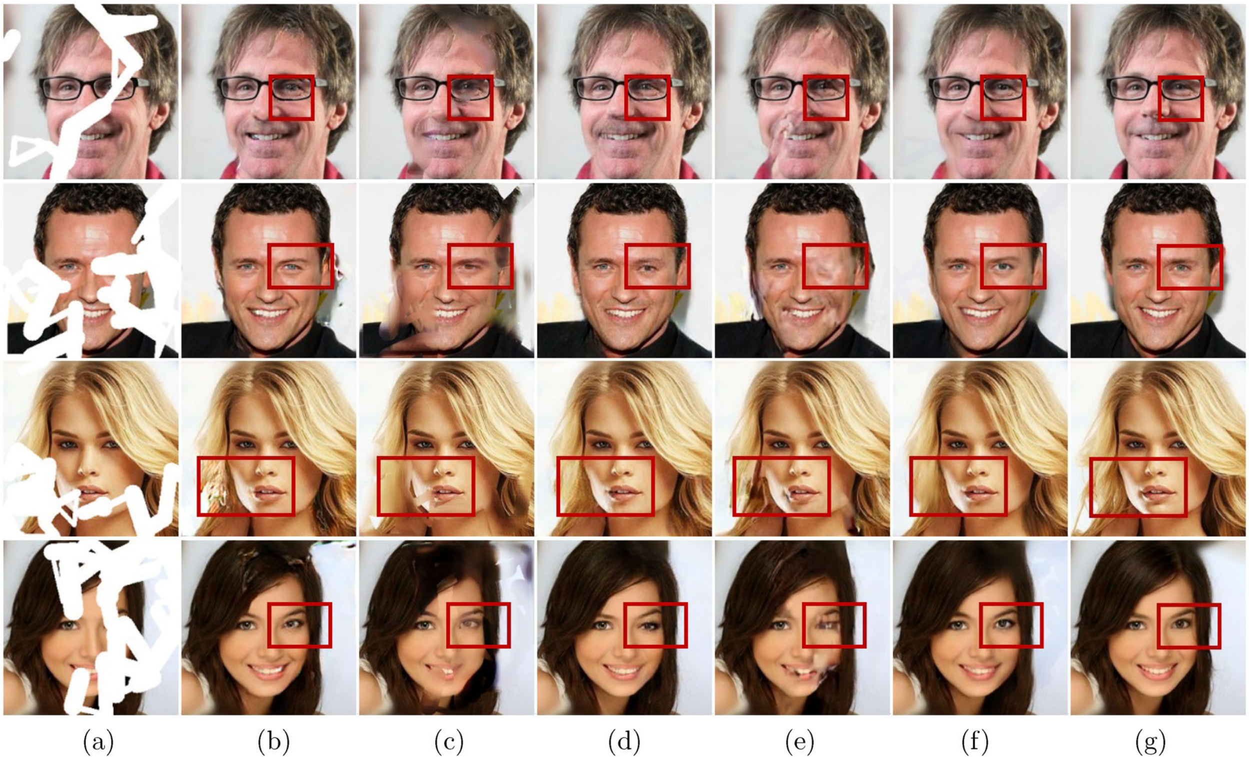


Fig. 5 Qualitative comparisons on CelebA-HQ dataset (zoom in for a better view). In each row, images from left to right: **a** Masked inputs, **b** MC [36], **c** PEN-Net [37], **d** EC [19], **e** GDS [40], **f** SG-GAN, **g** Ground truth

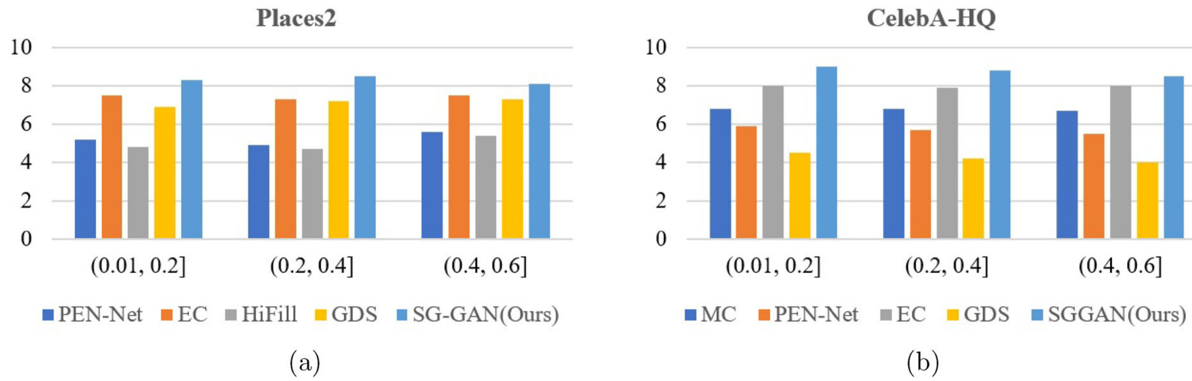


Fig. 6 User study results

Table 4 PSNR, SSIM, FID, and Mean L1 Loss for various completion networks on irregular masked images. The best result of each column is boldfaced. ↑ and ↓ represent larger and smaller as better, respectively

Method	PSNR↑	SSIM↑	FID↓	Mean L1 Loss↓
CA [1]	22.87	0.877	47.31	0.0652
PEN-Net [37]	22.17	0.870	54.10	0.0713
(A) Traditional conv	24.59	0.896	36.26	0.0526
(B) +Self attention	24.65	0.898	33.30	0.0508
(C) +CrA	**24.91**	**0.899**	**32.30**	**0.0501**

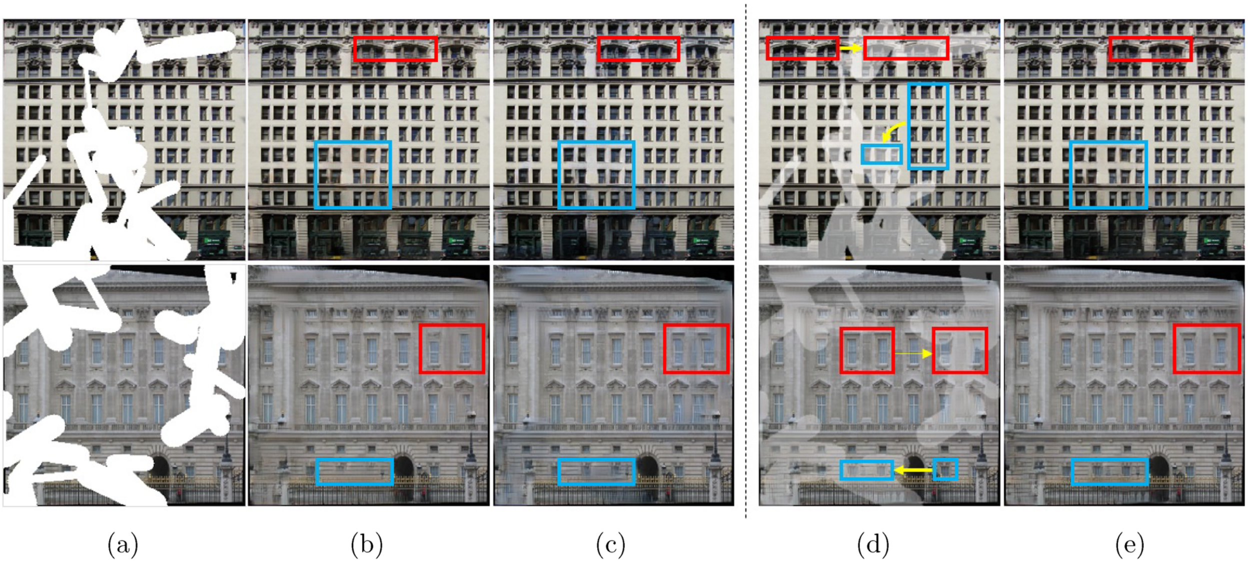


Fig. 7 Comparison results using CrA and Self Attention. **a** Masked inputs. **b** Results using Traditional Conv. **c** Results using Self Attention. **d** Flows from background regions to missing regions. **e** Results with CAC block

exhibit visual blurring when missing regions has complex structures (such as a green background in the third row). In addition, results from the first row in Fig. 4 also demonstrate that the SG-GAN method is superior in consistent structure restoration. Results obtained using different algorithms on CelebA-HQ dataset are shown in Fig. 5. It can be seen that, although MC [36] produces structurally reasonable results, there are artifacts such as water ripples in the hair. Results generated by PEN-Net [37] and GDS [40] tend to be blurry. The structures generated by EC [19] are inconsistent with the background, especially the eyes. The images generated by SG-GAN have clear facial features and better consistency with undamaged regions compared to the other baselines.

4.5 User study

A user study is further conducted to evaluate the subjective quality (Fig. 6). Specifically, 50 images are randomly

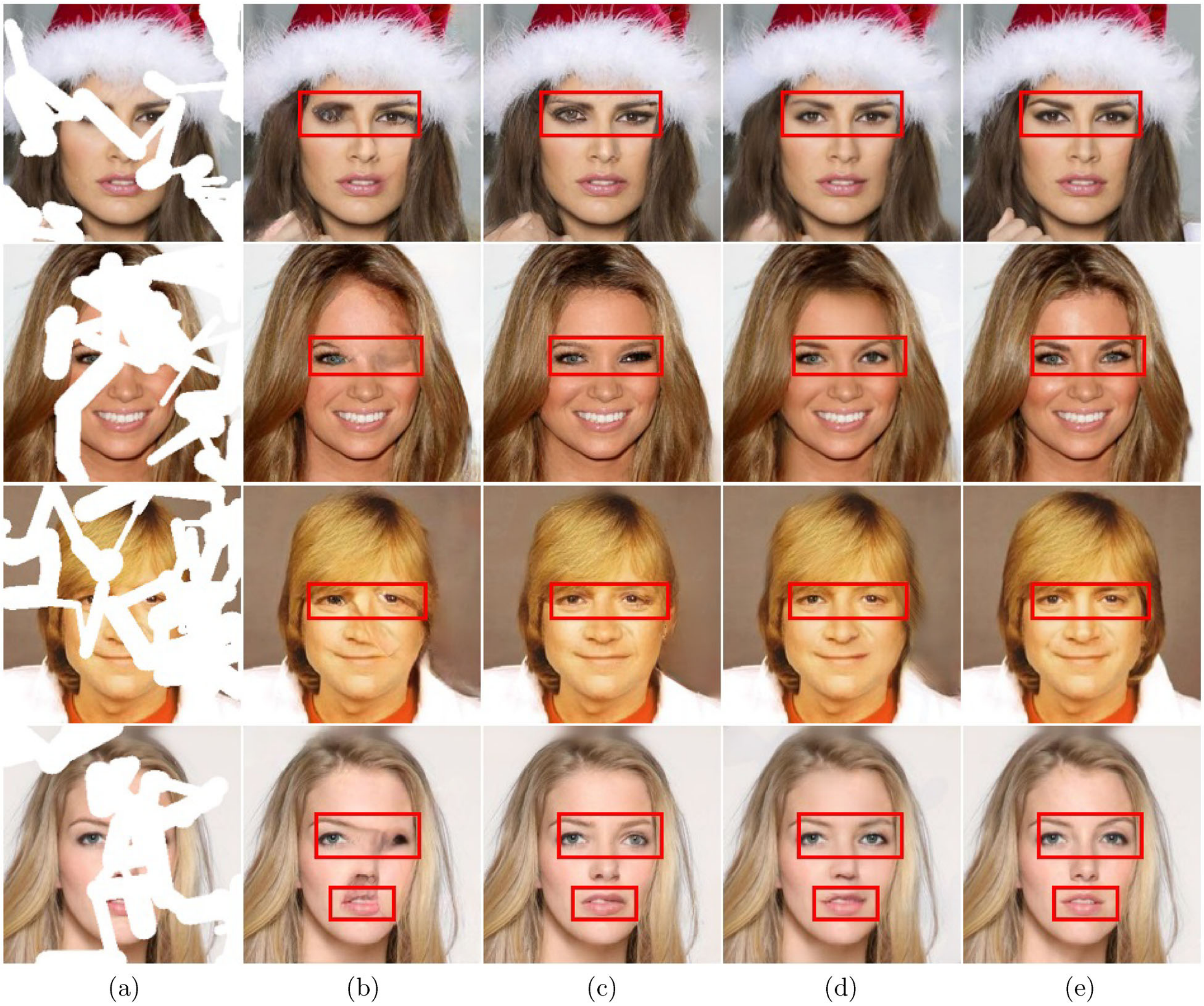


Fig. 8 Comparison results on CelebA-HQ dataset. In each row, images from left to right: **a** Masked inputs, **b** AOT [52], **c** TFill [28], **d** Results using CAC block, **e** Ground truth

selected from the Places2 test set and the CelebA-HQ test set, respectively. For each test image, different baseline methods are tested to obtain reconstruction results on mask ratios (0.01, 0.2], (0.2, 0.4], and (0.4, 0.6]. 25 participants are asked to rate the results from the highest image naturalness to the lowest image naturalness, with a score range of 0–10. Answers from 25 participants were collected and calculated the average score. The proposed SG-GAN has significant advantages under different mask ratios as shown in Fig. 6.

4.6 Ablation study

Several variants of the SG-GAN framework are conducted and ablation experiments are implemented to figure out the validity of components in the proposed method.

4.6.1 Traditional conv vs. CAC

The components in the redesigned generator architecture are first evaluated. Results show that the proposed architecture considerably improves performance as shown in Table 4. The baseline (A) uses an encoder-decoder structure derived from CA [1], which is a pure CNN baseline. Metric scores of Table 4 demonstrate that baseline (A) is comparable to CNN-based CA [1] and PEN-Net [37]. Self-attention module is added to the decoder in baseline (B), where all three inputs come from previous layer features, improving performance compared to baseline (A). Results of baseline (B) are shown in Fig. 7c. When the CrA layer proposed in baseline (C) is added to form a CAC block, the performance is significantly improved, as shown in Fig. 7e. It can generate more natural results that are closer to the ground truth compared with baseline (B), achieving a smooth transition from undamaged

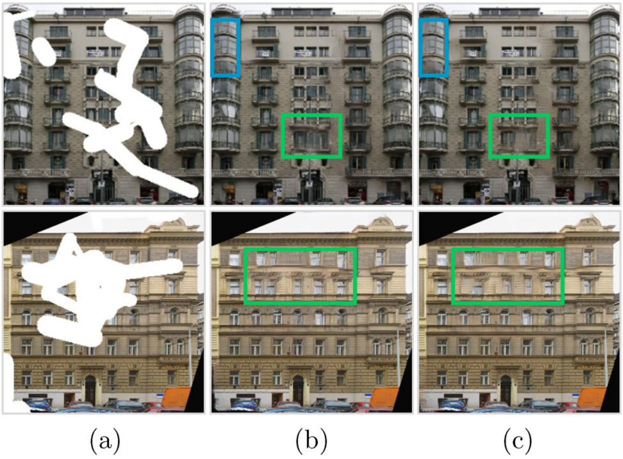


Fig. 9 Visual comparison of results using HRFM loss and not with it. **a** Masked inputs. **b** Results using HRFM loss. **c** Results without using HRFM loss

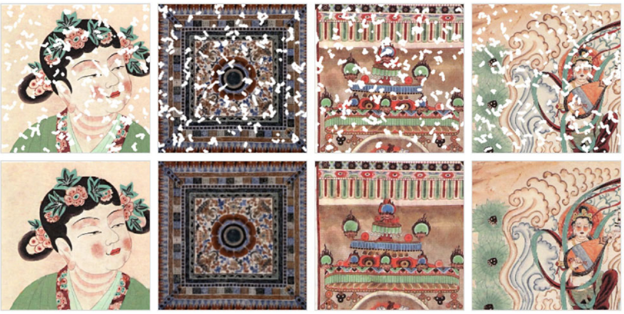


Fig. 10 Results on the Dunhuang Mural dataset. The first row is masked inputs and the second row is corresponding outputs

regions to missing regions, as shown in Fig. 7d. In addition, visual comparison results with TFill [28] and AOT [52] are presented on the CelebA-HQ dataset. Figure 8 demonstrates that the CAC module combining traditional convolution and CrA generates more consistent results than either pure CNN networks or transformer-based completion methods.

4.6.2 Analysis of HRFM loss

The effectiveness of HRFM loss is then verified in this section. A complete SG-GAN model with all losses and a model without HRFM loss are trained on the Façade dataset. The areas marked by blue and green rectangles in Fig. 9 demonstrate that the results generated by the model with all losses produce sharper edges than the network without HRFM loss on structure restoration.

4.7 Application

The proposed SG-GAN is also extended to Dunhuang mural [31] image restoration tasks in addition to performance verification on datasets Places2 and CelebA-HQ in Sects. 4.2 and 4.3. This application study is implemented on the Mural dataset. Here, block masks closer to real damage are generated to simulate the damaged region of mural images. The results are shown in Fig. 10. It can be seen that the proposed SG-GAN is able to generate visually complete and semantically reasonable results of Dunhuang mural images.

5 Conclusion

In this manuscript, an image inpainting method SG-GAN is first proposed, which recovers damaged images by a fusion learning paradigm of cross-attention and convolution. Secondly, an HRF subnet is constructed and an HRFM loss is proposed to enhance the understanding of the global structure and ensure the structural clarity of generated results. Extensive experiments demonstrate show that the proposed SG-GAN model is competent for image inpainting tasks and outperforms the state-of-the-art methods. Additionally, the proposed model has potential for extension to other tasks, e.g., image dereflection, which we will explore in the future work.

Acknowledgements This work was supported in part by the National Nature Science Foundation of China under Grant 62261032, and the Key Talent Project of Gansu Province.

Author Contributions CS and WL wrote the main manuscript text and prepared the figures. CS conducted the analysis of CAC block. CS, JM, and XJ performed deep learning experiments. WL and JL contributed to writing the manuscript. All authors reviewed the manuscript.

Availability of data and material The data that support the findings of this study are openly available in https://github.com/IPCSRG/SG-GAN_Inpainting.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T. S.: Generative Image Inpainting with Contextual Attention. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5505–5514 (2018). https://doi.org/10.1109/CVPR.2018.00577
2. Zhou, Y., Chen, Z., Li, P., Song, H., Chen, C.L.P., Sheng, B.: FSAD-Net: feedback spatial attention dehazing network. IEEE Trans. Neural Netw. Learn. Syst. **34**(10), 7719–7733 (2023). https://doi.org/10.1109/TNNLS.2022.3146004
3. Chen, Z., Qiu, G., Li, P., Zhu, L., Yang, X., Sheng, B.: MNGNAS: distilling adaptive combination of multiple searched networks for one-shot neural architecture search. IEEE Trans. Pattern Anal. Mach. Intell. **45**(11), 13489–13508 (2023). https://doi.org/10.1109/TPAMI.2023.3293885

4. Che, A., Yang, J.-H., Guo, C., Dai, H.-N., Xie, H., Li, P.: AEGAN: generating imperceptible face synthesis via autoencoder-based generative adversarial network. Comput. Animat. Virtual Worlds (2023). https://doi.org/10.1002/cav.2160
5. Li, P., Sheng, B., Chen, C.L.P.: Face sketch synthesis using regularized broad learning system. IEEE Trans. Neural Netw. Learn. Syst. **33**(10), 5346–5360 (2022). https://doi.org/10.1109/TNNLS.2021.3070463
6. Jiang, N., Sheng, B., Li, P., Lee, T.-Y.: PhotoHelper: portrait photographing guidance via deep feature retrieval and fusion. IEEE Trans. Multimed. **25**, 2226–2238 (2023). https://doi.org/10.1109/TMM.2022.3144890
7. Ma, X., Zhou, X., Huang, H., Jia, G., Chai, Z., Wei, X.: Contrastive attention network with dense field estimation for face completion. Pattern Recognit. (2022). https://doi.org/10.1016/j.patcog.2021.108465
8. Yu, T., Lin, C., Zhang, S., Wang, C., Ding, X., An, H., Liu, X., Qu, T., Wan, L., You, S., Wu, J., Zhang, J.: Artificial intelligence for Dunhuang cultural heritage protection: the project and the dataset. Int. J. Comput. Vis. **130**(11), 2646–2673 (2022). https://doi.org/10.1007/s11263-022-01665-x
9. Wang, N., Wang, W., Hu, W., Fenster, A., Li, S.: Thanka mural inpainting based on multi-scale adaptive partial convolution and stroke-like mask. IEEE Trans. Image Process. **30**, 3720–3733 (2021). https://doi.org/10.1109/TIP.2021.3064268
10. Sheng, B., Li, P., Jin, Y., Tan, P., Lee, T.-Y.: Intrinsic image decomposition with step and drift shading separation. IEEE Trans. Vis. Comput. Graph. **26**(2), 1332–1346 (2020). https://doi.org/10.1109/TVCG.2018.2869326
11. Sun, J., Yuan, L., Jia, J., Shum, H.-Y.: Image completion with structure propagation. ACM Trans. Graph. **24**(3), 861–868 (2005). https://doi.org/10.1145/1073204.1073274
12. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: PatchMatch: a randomized correspondence algorithm for structural image editing. ACM Trans. Graph. **28**(3), 24 (2009). https://doi.org/10.1145/1531326.1531330
13. Chen, Z., Zhou, Y., Li, R., Li, P., Sheng, B.: SCPA-Net: self-calibrated pyramid aggregation for image dehazing. Comput. Animat. Virtual Worlds (2022). https://doi.org/10.1002/cav.2061
14. Guo, X., Yang, H., Huang, D.: Image Inpainting via Conditional Texture and Structure Dual Generation. In: IEEE International Conference on Computer Vision, pp. 14114–14123 (2021). https://doi.org/10.1109/ICCV48922.2021.01387
15. Li, H., Sheng, B., Li, P., Ali, R., Chen, C.L.P.: Globally and locally semantic colorization via exemplar-based broad-GAN. IEEE Trans. Image Process. **30**, 8526–8539 (2021). https://doi.org/10.1109/TIP.2021.3117061
16. Guo, H., Sheng, B., Li, P., Chen, C.L.P.: Multiview high dynamic range image synthesis using fuzzy broad learning system. IEEE Trans. Cybern. **51**(5), 2735–2747 (2021). https://doi.org/10.1109/TCYB.2019.2934823
17. Sheng, B., Li, P., Fang, X., Tan, P., Wu, E.: Depth-aware motion deblurring using loopy belief propagation. IEEE Trans. Circuits Syst. Video Technol. **30**(4), 955–969 (2020). https://doi.org/10.1109/TCSVT.2019.2901629
18. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A. A.: Context Encoders: Feature Learning by Inpainting. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2536–2544 (2016). https://doi.org/10.1109/CVPR.2016.278
19. Nazeri, K., Ng, E., Joseph, T., Qureshi, F. Z., Ebrahimi, M.: EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning. In: IEEE International Conference on Computer Vision, pp. 3265–3274 (2019). arXiv:1901.00212
20. Ren, Y., Yu, X., Zhang, R., Li, T. H., Liu, S., Li, G.: StructureFlow: Image Inpainting via Structure-aware Appearance Flow. In: IEEE International Conference on Computer Vision, pp. 181–190 (2019). https://doi.org/10.1109/ICCV.2019.00027
21. Kim, J., Kim, W., Oh, H., Lee, S.: Progressive contextual aggregation empowered by Pixel-Wise confidence scoring for image inpainting. IEEE Trans. Image Process. **32**, 1200–1214 (2023). https://doi.org/10.1109/TIP.2023.3238317
22. Liu, G., Dundar, A., Shih, K.J., Wang, T.-C., Reda, F.A., Sapra, K., Yu, Z., Yang, X., Tao, A., Catanzaro, B.: Partial convolution for padding, inpainting, and image synthesis. IEEE Trans. Pattern Anal. Mach. Intell. **45**(5), 6096–6110 (2023). https://doi.org/10.1109/TPAMI.2022.3209702
23. Liu, W., Cao, C., Liu, J., Ren, C., Wei, Y., Guo, H.: Fine-grained image inpainting with scale-enhanced generative adversarial network. Pattern Recognit. Lett. **143**, 81–87 (2021). https://doi.org/10.1016/j.patrec.2020.12.008
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 6000–6010 (2017). arXiv:1706.03762
25. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-Attention Generative Adversarial Networks. In: International Conference on Machine Learning, (2019). arXiv:1805.08318
26. Dong, Q., Cao, C., Fu, Y.: Incremental Transformer Structure Enhanced Image Inpainting with Masking Positional Encoding. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 11358–11368 (2022). https://doi.org/10.1109/CVPR52688.2022.01107
27. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for image recognition at scale. In: International Conference on Learning Representations, (2021). arXiv:2010.11929
28. Zheng, C., Cham, T.-J., Cai, J., Phung, D.: Bridging Global Context Interactions for High-Fidelity Image Completion. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 11502–11512 (2022). https://doi.org/10.1109/CVPR52688.2022.01122
29. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: a 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**(6), 1452–1464 (2018). https://doi.org/10.1109/TPAMI.2017.2723009
30. Tylecek, R. Sara, R.: Spatial Pattern Templates for Recognition of Objects with Regular Structure. In: German Conference on Pattern Recognition, pp. 364–374 (2013). https://doi.org/10.1007/978-3-642-40602-7_39
31. Li, L., Zou, Q., Zhang, F., Yu, H., Chen, L., Song, C., Huang, X., Wang, X.: Line Drawing Guided Progressive Inpainting of Mural Damages (2022). arXiv:2211.06649
32. Liu, G., Reda, F. A., Shih, K. J., Wang, T.-C., Tao, A., Catanzaro, B.: Image Inpainting for Irregular Holes Using Partial Convolutions. In: European Conference on Computer Vision, pp. 89–105 (2018). https://doi.org/10.1007/978-3-030-01252-6_6
33. Yi, Z., Tang, Q., Azizi, S., Jang, D., Xu, Z.: Contextual Residual Aggregation for Ultra High-Resolution Image Inpainting. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 7508–7517 (2020). arXiv2005.09704
34. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust Large Mask Inpainting with Fourier Convolutions. In: IEEE Winter Conference on Applications of Computer Vision, pp. 3172–3182 (2022). https://doi.org/10.1109/WACV51458.2022.00323
35. Yang, Y., Cheng, Z., Yu, H., Zhang, Y., Cheng, X., Zhang, Z., Xie, G.: MSE-Net: generative image inpainting with multi-scale encoder. Vis. Comput. **38**(8), 2647–2659 (2022). https://doi.org/10.1007/s00371-021-02143-0

36. Wang, Y., Tao, X., Qi, X., Shen, X., Jia, J.: Image Inpainting via Generative Multi-column Convolutional Neural Networks. In: Advances in Neural Information Processing Systems, pp. 331–340 (2018). arXiv1810.08771
37. Zeng, Y., Fu, J., Chao, H., Guo, B.: Learning pyramid-context encoder network for high-quality image inpainting. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1486–1494 (2019). https://doi.org/10.1109/CVPR.2019.00158
38. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.: Free-form image inpainting with gated convolution. In: IEEE International Conference on Computer Vision, pp. 4471–4480 (2019). https://doi.org/10.1109/ICCV.2019.00457
39. Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E. I.-C., Xu, Y.: Large Scale Image Completion via Co-Modulated Generative Adversarial Networks. In: International Conference on Learning Representations, (2021). arXiv:2103.10428
40. Peng, J., Liu, D., Xu, S., Li, H.: Generating Diverse Structure for Image Inpainting With Hierarchical VQ-VAE. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 10775–10784 (2021). arXiv:2103.10022
41. Wang, T., Ouyang, H., Chen, Q.: Image Inpainting with External-internal Learning and Monochromic Bottleneck. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5116–5125 (2021). https://doi.org/10.1109/CVPR46437.2021.00508
42. Zhu, M., He, D., Li, X., Li, C., Li, F., Liu, X., Ding, E.: Image inpainting by end-to-end cascaded refinement with mask awareness. IEEE Trans. Image Process. **30**, 4855–4866 (2021). https://doi.org/10.1109/TIP.2021.3076310
43. Li, X., Guo, Q., Lin, D., Li, P., Feng, W., Wang, S.: MISF:Multi-level Interactive Siamese Filtering for High-Fidelity Image Inpainting. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1859–1868 (2022). https://doi.org/10.1109/CVPR52688.2022.00191
44. Iizuka, S., Serra, E.S., Ishikawa, H.: Globally and locally consistent image completion. ACM Trans. Graph. **36**(4), 14 (2017). https://doi.org/10.1145/3072959.3073659
45. Xie, Z., Zhang, W., Sheng, B., Li, P., Chen, C.L.P.: BaGFN: broad attentive graph fusion network for high-order feature interactions. IEEE Trans. Neural Netw. Learn. Syst. **34**(8), 4499–4513 (2023). https://doi.org/10.1109/TNNLS.2021.3116209
46. Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond. In: IEEE International Conference on Computer Vision, pp. 1971–1980 (2019). https://doi.org/10.1109/ICCVW.2019.00246
47. Fu, J., Liu, J., Tian, H., Li, Y.: Dual Attention Network for Scene Segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3141–3149 (2019). https://doi.org/10.1109/CVPR.2019.00326
48. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer-Hierarchical Vision Transformer using Shifted Windows. In: International Conference on Computer Vision, pp. 9992–10002 (2021). https://doi.org/10.1109/ICCV48922.2021.00986
49. Lin, X., Sun, S., Huang, W., Sheng, B., Li, P., Feng, D.D.: EAPT: efficient attention pyramid transformer for image processing. IEEE Trans. Multimed. **25**, 50–61 (2023). https://doi.org/10.1109/TMM.2021.3120873
50. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked Autoencoders Are Scalable Vision Learners. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 15979-15988 (2022). https://doi.org/10.1109/CVPR52688.2022.01553
51. Li, J., Chen, J., Sheng, B., Li, P., Yang, P., Feng, D.D., Qi, J.: Automatic detection and classification system of domesticwaste via multimodel cascaded convolutional neural network. IEEE Trans. Industr. Inform. **18**(1), 163–173 (2022). https://doi.org/10.1109/TII.2021.3085669
52. Zeng, Y., Fu, J., Chao, H., Guo, B.: Aggregated contextual transformations for high-resolution image inpainting. IEEE Trans. Vis. Comput. Graph. **29**(7), 3266–3280 (2023). https://doi.org/10.1109/TVCG.2022.3156949

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Changhong Shi received the M.S. degree from Lanzhou University of Technology, Lanzhou, China, in 2018. She is currently pursuing the Ph.D. degree in pattern recognition and intelligent systems at Lanzhou University of Technology, Lanzhou, China. Her main research interests are image processing and pattern recognition.

Weirong Liu received the M.S. degree from Lanzhou University of Technology, Lanzhou, China, and the Ph.D. degree, from Hunan University, Changsha, China. He is currently a professor with Lanzhou University of Technology, Lanzhou, China. His principal research interests are on topics related to image processing, pattern recognition and intelligent systems, and control theory and control engineering.

Jiahao Meng received his M.S. degree in Northwest Minzu University. He is currently pursuing the Ph.D. degree at Lanzhou University of Technology, Lanzhou, China. His main research interests are image processing and pattern recognition.

Xiongfei Jia is currently pursuing the M.S. degree at Lanzhou University of Technology, Lanzhou, China. He received his B.S. degree in Lanzhou University of Technology. His main research interests are image processing and pattern recognition.

Jie Liu received her M.S. degree in systems engineering from Gansu University of Technology, Lanz hou, China. She is currently an associate professor at Lanzhou University of Technology, Lanzhou, China. Her principal research interests are on topics related to image processing, pattern recognition, and intelligent systems.

Sensing and Imaging (2025) 26:98
https://doi.org/10.1007/s11220-025-00624-3

RESEARCH

Cultural Relic Image Inpainting via Multi-column Condition Decoding Transformer

Changhong Shi[1] · Weirong Liu[1] · Zhijun Li[1] · Jiajing Yi[1] · Jie Liu[1]

Received: 24 April 2025 / Revised: 6 June 2025 / Accepted: 16 June 2025

Abstract

Current cultural relic image inpainting methods mainly utilize single encoder-decoder architectures. However, single encoder-decoder methods struggle with introducing prior conditions, especially for historical relics datasets with unique damage patterns. A Multi-column Condition Decoding Transformer for Cultural Relic Image Inpainting (MCDT) is proposed to address above issue. The proposed MCDT model employs multi-column decoders integrated into the Transformer through cross-attention mechanism. Specifically, multi-column decoders consist of three branches: (1) a self-attention branch that decodes the encoded latent features, (2) a ground-truth cross-attention branch that enforces constraints from ground truth data, and (3) an edge cross-attention branch that incorporates edge constraints. The multi-column decoding architecture enables the simultaneous integration of multiple external conditions to constitute a multi-prior constrained image inpainting model. Comparative experiments conducted on cultural relic dataset show that the proposed MCDT method generates higher-quality inpainting results compared to state-of-the-art methods.

Keywords Cultural relic image inpainting · Transformer · Cross attention · Multi-column decoder

✉ Weirong Liu
liuwr@lut.edu.cn

[1] College of Electric and Informational Engineering, Lanzhou University of Technology, LanZhou, Gansu 730050, China

Published online: 30 June 2025

1 Introduction

Image inpainting, also referred to as image completion, is a computer vision task that reconstructs or infers missing regions using information from undamaged regions. The objective of image inpainting task is to synthesize semantically coherent and visually realistic image content, which has been applied in multiple fields including cultural heritage preservation, forensic analysis, film production, and historical photograph restoration [1]. Image inpainting methods can be categorized into three distinct categories: traditional approaches, generative adversarial network (GAN)-based methods, and Transformer-based techniques.

A “copy-and-paste” thought is employed to transfer features from undamaged regions to missing ones in traditional image inpainting methods. However, due to the lack of comprehension regarding high-level semantic information, traditional methods are observed to fail in generating plausible content.

The adversarial mechanism of GAN network [2] has been widely adopted in autoencoder-based image inpainting methodologies [3–11]. Autoencoder-based image inpainting approaches employ a generative network and a discriminative network. They collaboratively learn to synthesize missing regions through competitive interaction. The autoencoder-based approaches can handle large missing regions and improves both detail richness and semantic consistency. Pathak et al. [6]. integrate the encoder-decoder structure with GANs, and propose the Context Encoders CE framework, exhibiting its global image understanding and restoration capabilities. Various improved autoencoder-GAN models have been subsequently developed following by CE, including the Global-Local GL consistency model [5], the Contextual Attention CA composite inpainting model [4], the Generative Multi-column CNN GMCNN [7], the Pyramid Context Encoder Network PEN-Net [10], the EdgeConnect EC inpainting model [9], the progressive hierarchical GAN inpainting network [8], and the dense discriminative model proposed by Cao et al. [12]. However, autoencoder-based GAN methods exhibit detail loss during feature compression encoding, and the detail loss often results in overly smoothed restoration outputs with blurred boundaries.

Transformer-based image inpainting approaches leverage the inherent self-attention mechanisms of Transformer architectures to model long-range dependencies and global contextual relationships among different image regions [13–15]. Transformer-based image inpainting approaches achieve image inpainting entirely through attention-based operations, eliminating conventional convolutional layers. The core self-attention mechanism facilitates comprehensive spatial relationship modeling. Dosovitskiy et al. [16]. first propose the Vision Transformer (ViT) architecture to apply Transformer to image recognition, promoting superior performance compared to CNN-based image inpainting methods. Zheng et al. [13]. propose a hybrid architecture that combines Transformers and CNNs to leverage the complementary advantages of both frameworks. The study reveals that hybrid methods exhibit inconsistent convergence between convolution and self-attention layers during joint training, which frequently leads to training instability. Li et al. [14]. propose the MAT model by integrating StyleGAN and Transformers. MAT applies the diverse generation capability in StyleGAN to produce varied inpainting results from identical damaged inputs and introduces diversity in image inpainting tasks. Concurrently, diffusion Transformer models that share conceptual similarities with StyleTrans approaches [17] have emerged [15]. Diffusion Transformer models apply Markov-based forward-reverse diffusion mechanisms. The mechanisms introduce stochasticity in image inpainting processes.

Nevertheless, most ViT-based methods only utilize the self-attention mechanism which is one of the two critical modules in Transformers, while neglecting the equally essential module cross attention. ViT-based methods cannot incorporate external conditional priors such as ground truth or semantic information, leading to visually unnatural inpainting results. Therefore, it is difficult to be applied to image inpainting tasks with too few samples in real scenarios, such as the cultural relic image inpainting, etc. It becomes imperative to fully leverage limited datasets and employ models with robust modeling capabilities to achieve effective restoration of damaged cultural relics. The input of current Transformer architectures is constrained by self-decoding, lacking multimodal inputs as shown in Fig. 1. The cross attention mechanism is utilized in this manuscript to integrate structural and ground truth constraints into the decoding process to address above limitation, thereby transforming single-condition decoding into multi-condition decoding. The proposed approach maximizes the utilization of prior knowledge in data-scarce scenarios.

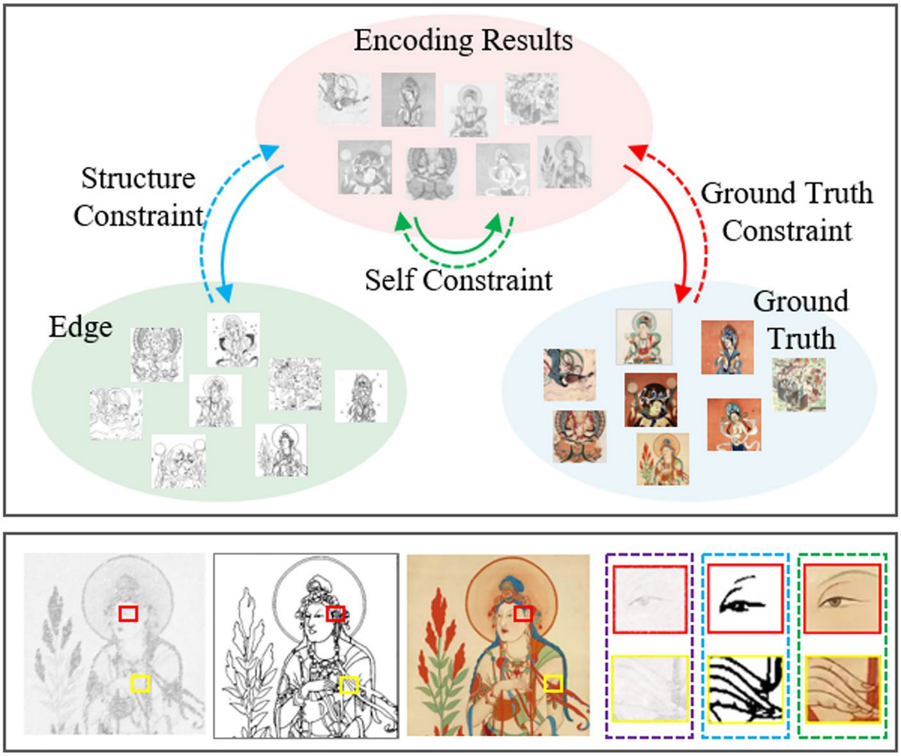


Fig. 1 Cross-input for decoding under multiple constraints

Recently, the latent diffusion model proposed by Rombach et al. innovatively employs cross attention to introduce external conditions, enabling the interaction of semantic, textual, and visual information with encoded noise. The latent diffusion model establishes a mapping relationship between noise and external information, significantly enhancing the comprehension capabilities. Inspired by the latent diffusion model, this manuscript leverages cross attention to incorporate external prior knowledge, thereby expanding the scope of information utilization. Specifically, a multi-column condition decoding transformer is proposed for the restoration of cultural relic images. External priors such as ground truth and structural information are introduced into the Transformer decoder by utilizing cross attention mechanism. These external priors are then subjected to cross attention operations with the encoded results, forming two distinct decoding branches: ground truth cross attention decoding and structural cross attention decoding. These two branches collectively constitute a multi-column condition decoder along with the self-attention decoding branch. Multi-column condition decoder captures essential long-range dependencies for image inpainting. Multi-column condition decoder models fine-grained dependencies connecting encoder outputs with ground truth data and establishes coarse-grained dependencies linking encoder features to structures. The proposed method expands the range of utilizable information, and generates visually authentic and structurally coherent results, by imposing constraints at both fine-grained and coarse-grained levels.

In a nutshell, the main contributions are summarized as follows:

1. A novel multi-column condition decoding transformer architecture is proposed for cultural relic image inpainting. Specifically, external prior knowledge is incorporated into the decoder through cross attention mechanisms. A triple-branch decoder consisting of self attention branch, ground truth cross attention branch and edge cross attention branch is constructed, enabling dual guidance from both fine-grained ground truth constraints and coarse-grained structural constraints.
2. Experiments on damaged cultural relic images are conducted, and the experiment results show that the proposed model achieves superior inpainting performance. The generated images are verified to possess clearer structures and more authentic textures compared with state-of-the-art methods.

2 Related Work

2.1 Image Inpainting on Cultural Relic Images

Existing mural image inpainting methods can be divided into traditional inpainting methods and deep learning inpainting methods. Although traditional methods [18] have relatively high computational efficiency, they struggle to restore complex damage patterns. mural restoration methods based on deep learning have been proposed successively to solve this problem. For example, Yu et al. [19] proposes an end-to-end partial convolution U-Net for Dunhuang wall-painting inpainting, which employs a hybrid loss function combining transition variation, content, and style losses. The end-to-end image inpainting model utilizes synthetic degraded data to reconstruct non-rigid, irregular missing regions while preserving the original artistic style. Subsequently, Yu et al. [20] presents an AI-powered project and associated dataset for Dunhuang cultural heritage preservation, focusing on digital inpainting, analysis, and conservation of murals and artifacts through intelligent algorithms. Zhou et al. [21] proposes a structure-guided deep network for Dunhuang mural inpainting, which integrates structural cues (e.g., lines, contours) with deep learning to achieve high-fidelity restoration. Zeng et al. [22] proposes an image inpainting method based on aggregated context transformation. It maintains the semantic consistency and detailed authenticity of the inpainting area by integrating multi-scale feature transformation and attention mechanisms. In this work, a novel Transformer-based image inpainting framework is developed to addresses prior-condition integration in cultural relic inpainting by jointly leveraging multiple constraints.

2.2 Transformer

Transformer is an encoder-decoder architecture composed of multiple learnable self-attention layers, the architecture diagram is shown in Fig. 2. Transformer has been successfully adapted for various visual tasks [23], including image classification, image generation, and image completion [1, 13, 24–27]. Transformer has over-

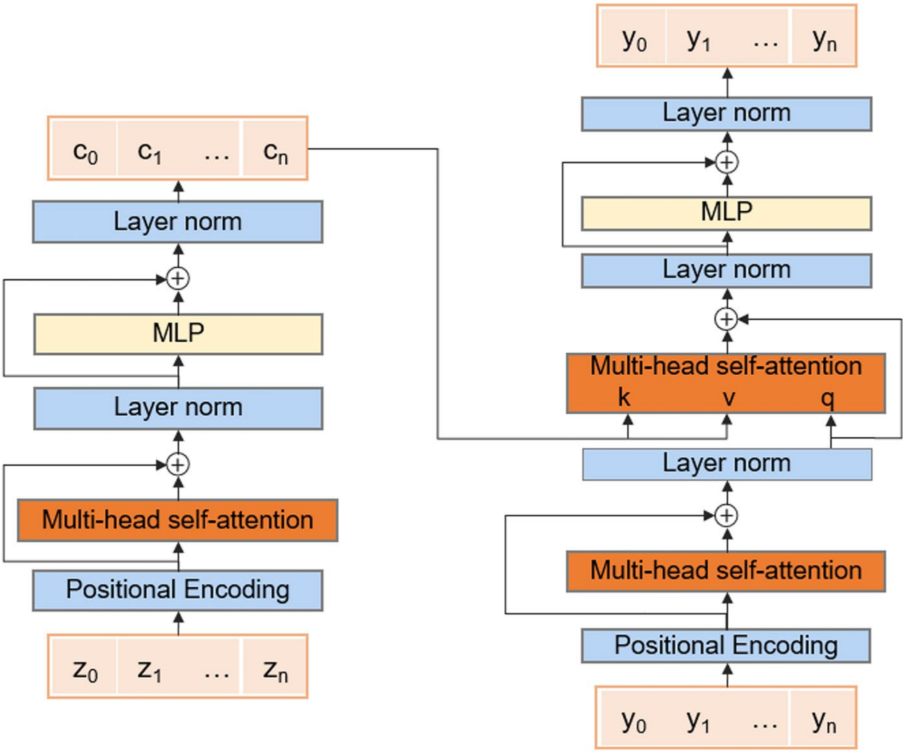


Fig. 2 Transformer architecture diagram

come the limitations of sequence dependency inherent in traditional RNNs and the restricted receptive fields of CNNs. Global dependencies among all elements in the input sequence are computed in parallel through multi-head attention mechanisms, while sequential order information is preserved through positional encoding. Both the encoder and decoder components are constructed using stacked attention layers and feed-forward networks, with residual connections and layer normalization being incorporated to ensure training stability. Transformer has been established as a foundational model in both natural language processing and computer vision fields owing to its exceptional capability in capturing long-range dependencies and supporting parallel computation.

3 Method

3.1 Network Architecture

A multi-column condition decoding Transformer model MCDT is proposed to enhance the constraint capability of single-decoder inpainting models. The framework is illustrated in Fig. 3. Each decoder branch is designed to impose constraints on different modalities: the self-attention branch directly reconstructs encoded features, the edge cross attention branch incorporates structural prior constraints, while the ground truth cross attention branch establishes detailed mapping between generated results and ground truth.

Specifically, the encoder comprises four Transformer encoder blocks, each constructed with self-attention layers. Following each encoding layer, downsampling operations are performed to progressively reduce feature dimensions, thereby enabling deep feature extraction and enhancing the multi-scale representation capability. The resulting encoded features are then processed by the decoder, which consists of three parallel branches, each containing three Transformer decoder blocks. The damaged input image I_M is transformed into latent features through the encoder, the process that can be formally expressed as:

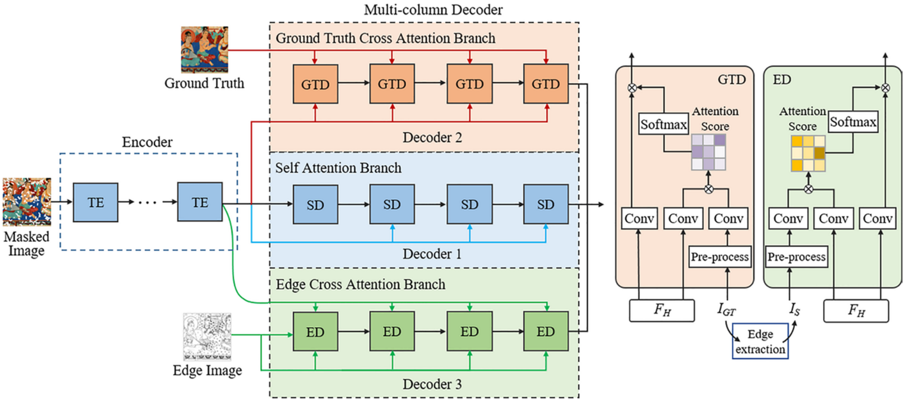


Fig. 3 The multi-column condition decoding transformer framework

$$F_H = T_e(I_M) \tag{1}$$

where the encoding mapping process is denoted by $T_e(\cdot)$, F_H represents the output features from the encoder.

Subsequently, the encoded results F_H are fed into the three parallel decoder branches to generate the final output, namely the restored image I_{OUT}. The reconstruction process can be expressed as:

$$I_{OUT} = T_{md}(F_H) \tag{2}$$

where $T_{md}(\cdot)$ denotes the decoding mapping process, I_{OUT} represents the output features from the multi-column condition decoder.

3.2 Self Attention Branch

The self-attention branch, structurally similar to the encoder, is composed of self-decoder (SD) blocks. SD block is to restore the encoded results F_H to the original image dimensions through inverse scaling transformations.

Taking the first SD block as an example, the computation process can be formulated as:

$$\begin{aligned} f_0 &= \boldsymbol{W}_{f0} \cdot Conv1(F_H) \\ g_0 &= \boldsymbol{W}_{g0} \cdot Conv1(F_H) \\ h_0 &= \boldsymbol{W}_{h0} \cdot Conv1(F_H) \end{aligned} \tag{3}$$

where $\boldsymbol{W}_{f0}$, $\boldsymbol{W}_{g0}$, and $\boldsymbol{W}_{h0}$ are learnable weight matrices. $Conv1(\cdot)$ is 1 × 1 convolution operation. The self attention weight s_0 can be calculated by softmax function:

$$s_0 = soft\ \max(f_0 {g_0}^{\mathrm{T}}) \tag{4}$$

3.3 Ground Truth Cross Attention Branch

Cross attention mechanisms can be employed to introduce additional valuable information into the model. A ground truth cross attention branch is proposed to jointly process both ground truth and hidden features during decoding. The ground truth cross attention branch is composed of Ground Truth Decoder (GTD) blocks, whose inputs consist of encoded features and preprocessed ground truth.

Firstly, the preprocessing operation $P(\cdot)$ includes feature extraction and downsampling procedures designed to transform the ground truth into features that share the same modality with the encoded structure F_H. And then, both F_H and the preprocessed ground truth are fed into 1 × 1 convolution layers, the GTD output is generated after cross attention computation. Taking the first GTD block as an example, the computation process can be formulated as:

$$\begin{aligned} f_1 &= \boldsymbol{W}_{f1} \cdot Conv1(P(I_{GT})) \\ g_1 &= \boldsymbol{W}_{g1} \cdot Conv1(F_H) \\ h_1 &= \boldsymbol{W}_{h1} \cdot Conv1(F_H) \end{aligned} \tag{5}$$

where, $\boldsymbol{W}_{f1}$, $\boldsymbol{W}_{g1}$, and $\boldsymbol{W}_{h1}$ are learnable weight matrices. $P(\cdot)$ is the pre-process operation.The ground truth cross attention weight s_1 can be calculated by softmax function:

$$s_1 = soft\ \max(f_1 {g_1}^{\mathrm{T}}) \tag{6}$$

3.4 Edge Cross Attention Branch

The cross-attention mechanism enables not only intra-modal processing of heterogeneous inputs but also inter-modal establishment of long-range dependencies. To leverage this property for introducing additional modal information and enforcing structural constraints on image inpainting, an edge cross attention branch is proposed to operate in parallel with both the self attention branch and the ground truth cross attention branch. The edge cross attention branch is implemented through Edge Decoder (ED) blocks. ED blocks receive two primary inputs, the encoded features and structural information I_s. I_s is extracted from the ground truth.

Firstly, the structural input is obtained by applying feature extraction operations to the ground truth image, the computation process can be formulated as:

$$I_S = P_e(I_{GT}) \tag{7}$$

where $P_e(\cdot)$ represents the structural feature extraction operation. And then, The structural input I_s and encoded results F_H are processed through 1 × 1 convolutional layers. The output of ED block is generated after cross attention computation. The computation procedure for the first ED block can be formulated as:

$$\begin{aligned} f_2 &= \boldsymbol{W}_{f2} \cdot Conv1(I_S) \\ g_2 &= \boldsymbol{W}_{g2} \cdot Conv1(F_H) \\ h_2 &= \boldsymbol{W}_{h2} \cdot Conv1(F_H) \end{aligned} \tag{8}$$

where $\boldsymbol{W}_{f2}$, $\boldsymbol{W}_{g2}$, and $\boldsymbol{W}_{h2}$ are learnable weight matrices. The edge cross attention weight s_2 can be calculated by softmax function:

$$s_2 = soft\ \max(f_2 {g_2}^{\mathrm{T}}) \tag{9}$$

Finally, the final reconstructed output I_{OUT} is obtained by aggregating the outputs from all three decoder branches:

$$I_{OUT} = Conc(F_{GT}, F_S, F_E) \tag{10}$$

where, F_{GT}, F_S, and F_E represent the outputs from the ground truth cross attention branch, self-attention branch, and edge cross attention branch respectively, $Conc(\cdot)$ denotes the aggregation operation.

3.5 Loss Function

The proposed framework MCDT is trained using a joint loss function that combines adversarial loss, reconstruction loss, and style loss, ensuring the generation of both visually realistic and semantically coherent results. The joint loss function is formulated as an additive combination of the constituent terms:

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda_r \mathcal{L}_{rec} + \lambda_s \mathcal{L}_{style} \tag{11}$$

where λ_r and λ_s are regularity factors to balance contributions of losses.

Reconstruction Loss aming at achieving pixel-level consistency between the generated results and ground truth, the reconstruction loss $\mathcal{L}_{rec}$ includes pixel-level L1 loss and VGG loss is utilized. L1 Loss $\mathcal{L}_1$ is obtained by calculating the similarity of pixels, and VGG loss $\mathcal{L}_{vgg}$constrains information by extracting deep features:

$$\mathcal{L}_{rec} = \|(I_{GT} - I_{OUT})\|_1 + \|VGG_d(I_{GT}) - VGG_d(I_{OUT})\|_1 \tag{12}$$

where $VGG_d(\cdot)$ is a pre-trained CNN network, *d* is the feature map of specific VGG layer.

Adversarial Loss Discriminator is trained for pushing generator to reach its goal by distinguishing in the synthesized image from ground truth, and the adversarial loss plays a key role in game process. The adversarial loss can be formulated as:

$$\mathcal{L}_{adv} = E_{I_{GT} \sim P_{data}}[\log D(I_{GT})] + E_{I_{OUT} \sim P_G}[\log(1 - D(I_{OUT}))] \tag{13}$$

where P_{data} and P_G denote the real data distribution and generated data distribution respectively, $D(\cdot)$ denotes discriminator network.

Style Loss The style loss is introduced to ensure style consistency. An auto-correlation (Gram matrix) on each specific VGG feature map is performed before applying L1 loss. The style loss can be defined as:

$$\mathcal{L}_{style} = \sum n \left\| (\psi_n(I_{OUT}))^T \psi_n(I_{OUT}) - \psi_n(I_{GT}))^T \psi_n(I_{GT}) \right\|_1 \tag{14}$$

where $\psi_n(\cdot)$ is the activation map of the *n*-th selected layer.

4 Experiments

Comprehensive experiments were conducted to evaluate the proposed MCDT model, including simulated damage image inpainting, real damage image inpainting, ablation study, application study, and user study.

4.1 Datasets and Compared Methods

The validation of MCDT model was performed on Dhmurals1714 dataset [28], where 100 images were randomly selected for testing while the remaining 1,614 images constituted the training set. The Dhmurals1714 dataset was expanded through random cropping, rotation, and flipping operations to augment the training data, yielding 12,385 images for training. Random masks were employed during training, while testing utilized two types of masks: simulated irregular masks and authentic mural damage masks. MCDT was compared against the following state-of-the-art methods:

CA [4]: A GAN-based two-stage (coarse-to-fine) image inpainting approach that employs contextual attention mechanisms to match and transfer similar features from undamaged regions to missing regions.

AOT [22]: A generative adversarial image inpainting network. This method achieves long-range context modeling through stacked multi-scale AOT blocks in the generator and combines the mask prediction task of the discriminator to drive texture generation.

W-Net [29]: A dual-branch interactive inpainting network that establishes structure-texture relationships through coordinated structure-aware and texture-generation branches integrated with cross-modality attention fusion modules.

4.2 Experimental Setup

The hardware platform employed for both training and testing of the proposed MCDT method consisted of an Intel(R) Core(TM) i7-8700 CPU (3.2 GHz), a single NVIDIA GeForce RTX 4090 GPU (24 GB). The system environment was configured with Windows 10, Python 3.7, PyTorch 1.7.1, CUDA v9.0, and cuDNN v7.0. A batch size of 5 images was utilized during the training phase. The Adam optimizer was employed for loss function optimization, with first-order (beta1) and second-order (beta2) momentum parameters set to 0 and 0.99 respectively, while the learning rate was initialized at 1e-3. the loss function balance parameters were determined as $\lambda_r = 1.2$ and $\lambda_s = 0.01$ through hyperparameter tuning. MCDT architecture was designed with L=4 Transformer submodules for both the encoder and each decoder column. Due to the incorporation of cross attention layers, the generator was operated in two distinct modes. Both ground truth and hidden layer features participated in self-attention layer learning in training phase, while only hidden layer features were fed into the trained generator in testing phase. Performance evaluation was conducted using three metrics Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) and Fréchet Inception Distance (FID).

4.3 Simulated Damage Image Inpainting

Quantitative comparative experiments were conducted to evaluate the proposed MCDT model. MCDT model achieved superior performance compared to most existing mainstream methods in terms of PSNR, SSIM, and FID metrics. Table 1 presents the quantitative evaluation results of different methods on Dhmurals1714 dataset, with mask ratios categorized into three ranges: (0.01, 0.2], (0.2, 0.4], and (0.4, 0.6]. The results in Table 1 indicate that MCDT almost achieved superior values across all metrics compared with state-of-the-art methods. Specifically, MCDT improved PSNR by an average of 15%, increased SSIM by 0.019, and reduced FID by 36% compared to the CA model for the mask ratio range (0.01, 0.2]. MCDT achieved 0.87dB higher PSNR, 5% better SSIM score, and 22% lower FID compared to the W-Net method. MCDT reduced the average FID by 0.58, final average is 43.55, improved PSNR by 0.21dB, and increased SSIM by 0.004 compared to the AOT method for the mask ratio range (0.4, 0.6]. The experiment results show that MCDT model achieved the best average PSNR, SSIM, and FID values among all benchmark comparison methods.

A comparative visualization of the inpainting results generated by the proposed MCDT framework and existing benchmark methods is presented in Fig. 4. W-Net can reduce artifacts compared to CA as shown in the first row, but it fails to reconstruct the pattern texture properly with significant detail loss in the repetitive pattern region. AOT produces structurally misalign-ed inpainting results although generating some apparent details. In contrast, the output generated by MCDT exhibits semantically reasonable content, seamless background transition, and sharper edges, showing significant visual improvement. Severe artifacts and blurring effects are observed in CA results for the third-row image in Fig. 4, while both W-Net and AOT methods exhibit structural discontinuities. The proposed MCDT approach achieves the best visual quality compared with the other three methods. Results generate by MCDT has more continuous edges, richer details, and closest resemblance to the ground truth image. MCDT generates clear stripe patterns that blend naturally with the background regarding the last-row image. Results of CA method suffer from noticeable blurring. Whereas W-Net and AOT, present discontinuous edges and visually unnatu-

Table 1 The quantitative comparison on Dhmurals1714 dataset. The best result of each column is boldfaced. ↑ and ↓ represent larger and smaller is better, respectively

Metrics	Methods	Mask ratio		
		(0.01, 0.2]	(0.2, 0.4]	(0.4, 0.6]
PSNR$^{\uparrow}$	CA	26.29	22.68	20.11
	W-Net	30.05	27.41	25.32
	AOT	30.89	28.13	26.46
	MCDT	**30.92**	**28.21**	**26.67**
SSIM$^{\uparrow}$	CA	0.943	0.902	0.865
	W-Net	0.952	0.904	0.907
	AOT	0.961	0.942	0.922
	MCDT	**0.962**	**0.945**	**0.926**
FID$^{\downarrow}$	CA	49.91	64.83	76.44
	W-Net	40.83	53.52	64.36
	AOT	**31.42**	39.84	44.13
	MCDT	31.91	**39.24**	**43.55**

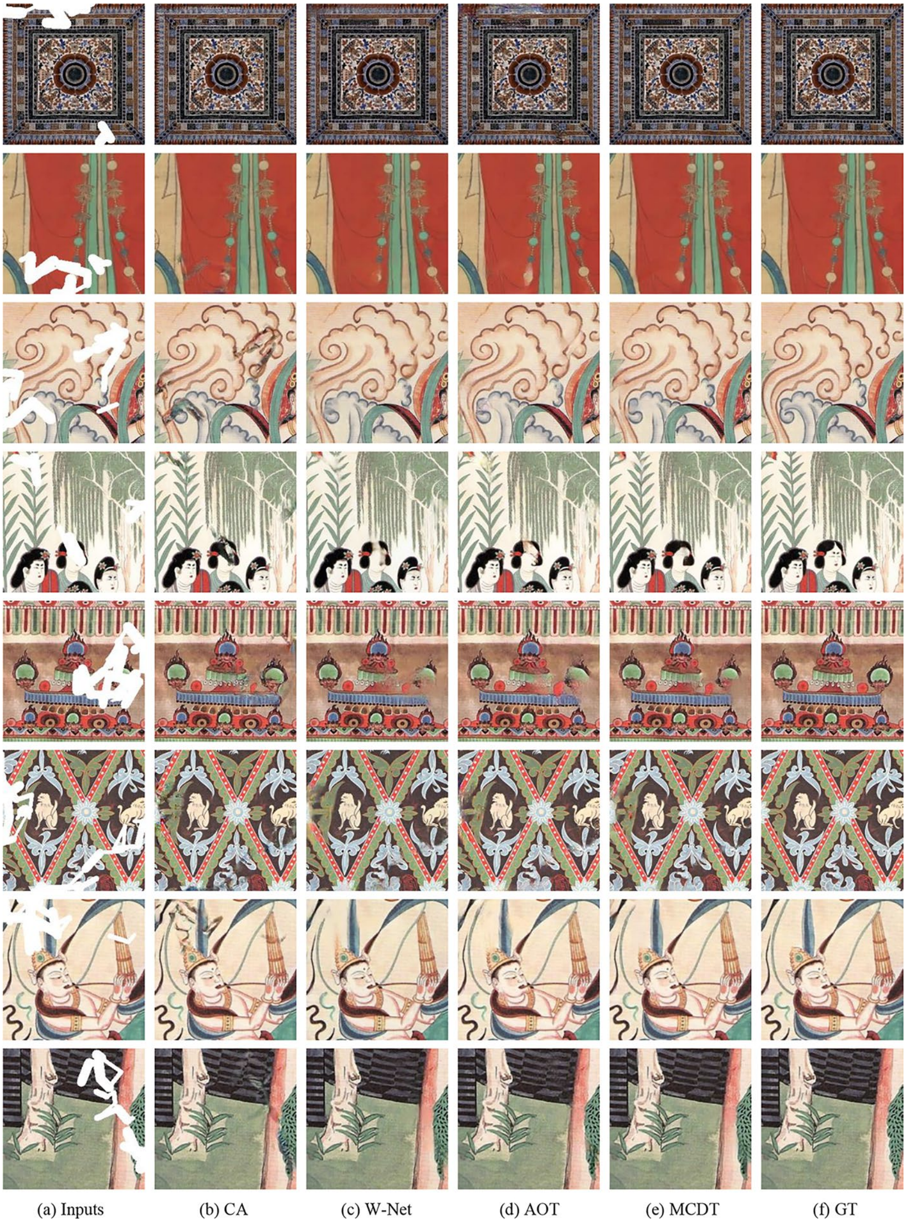


Fig. 4 Qualitative comparisons on simulation irregular mask (zoom in for a better view)

ral appearances while alleviating the blurring effect to some extent. Hence, the visual comparisons further show the superiority of MCDT model in terms of both measurement metrics and perceptual quality combined with the quantitative evaluation results.

4.4 Real Damage Image Inpainting

The proposed MCDT model was validated through the restoration of authentic damaged mural images. Specifically, this manuscript first performed precise annotation of the damaged regions in Fig. 5(f) to generate masks. Subsequently, the masked image in Fig. 5(a) was used as input for different inpainting algorithms. The output results from various algorithms are presented in Fig. 5. The results obtained by CA exhibit significant blurring artifacts as shown in the second and fifth rows of Fig. 5. While W-Net successfully avoids the matching errors characteristic of CA, its outputs are nevertheless compromised by severe pseudo-artifacts, as evidenced in the fourth and final rows. AOT suffers from over-inpainting issues though producing complete restorations, particularly noticeable in the reconstructed head region of the Buddha figure (fifth row). Conspicuous color abnormalities (third row) substantially degrade the overall visual quality. In marked contrast to the outputs generated by CA, W-Net and AOT, MCDT eliminates both blurring and distortion artifacts. The generated content of MCDT is semantically coherent, with smooth transitions and visual continuity between missing and background regions.

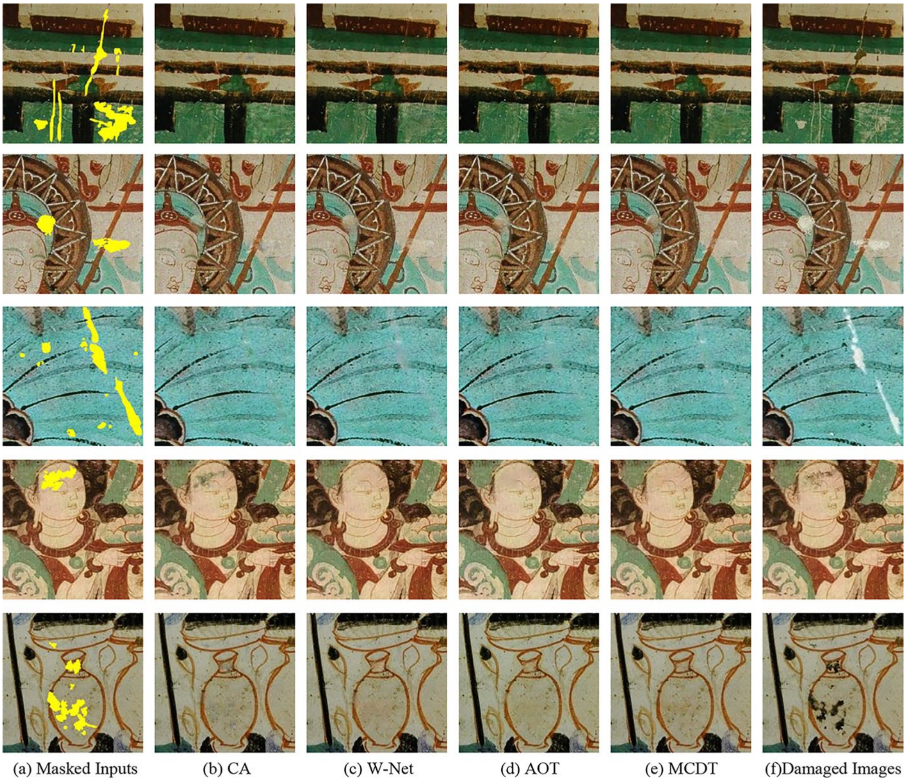


Fig. 5 Qualitative comparisons on images with real damage (zoom in for a better view)

4.5 Ablation Study

Key modules in MCDT framework were evaluated by ablation study, with particular focus on validating the effectiveness of cross attention mechanisms. The models were both trained and evaluated on Dhmurals1714 dataset. Two simplified model variants were first trained, one with the ground truth cross attention decoder (GTD) removed, and another with the edge cross attention decoder (ED) eliminated. The two simplified models were then visually compared against the complete MCDT model. Figure 6 (a) shows the input damaged image. Figure 6 (b) reveals that removal of GTD leads to noticeable edge blurring and detail loss. Figure 6 (c) indicates that the simplified model suffers from structural distortion in complex texture regions. Figure 6 (d) shows that the complete MCDT model maintains more accurate geometric structures and significantly outperforms other variants in detail richness and naturalness, particularly in reconstructing regular repetitive texture patterns that closely approximate the high-resolution reference image shown in Fig. 6 (e). Hence, the model MCDT achieves more accurate reconstruction of high-frequency details and sharper edges by establishing global dependencies between Transformer submodules and external priors.

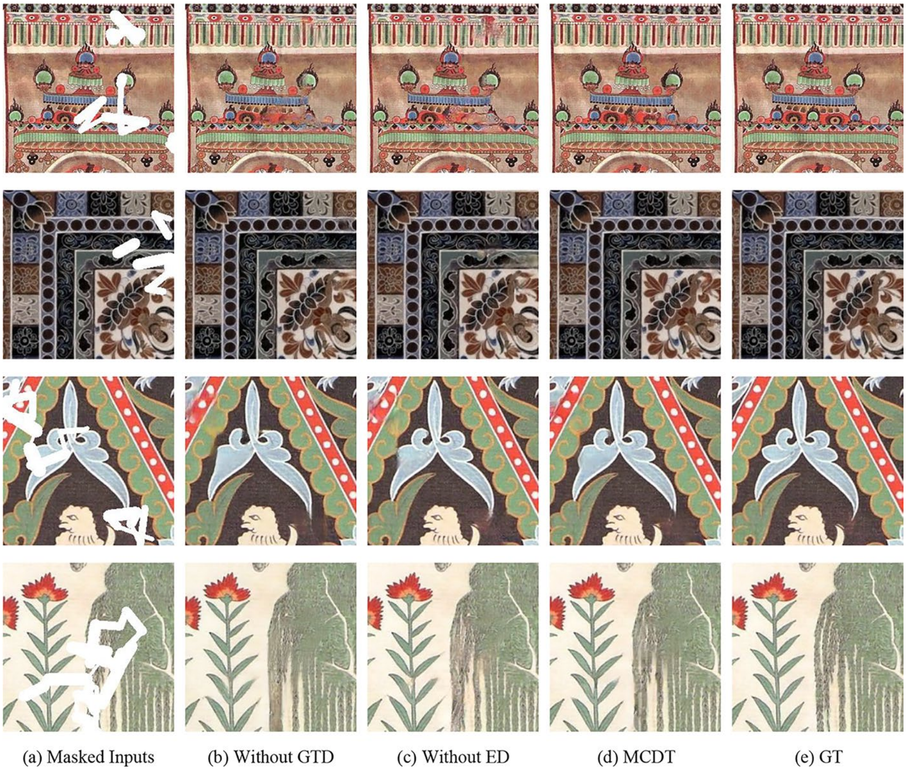
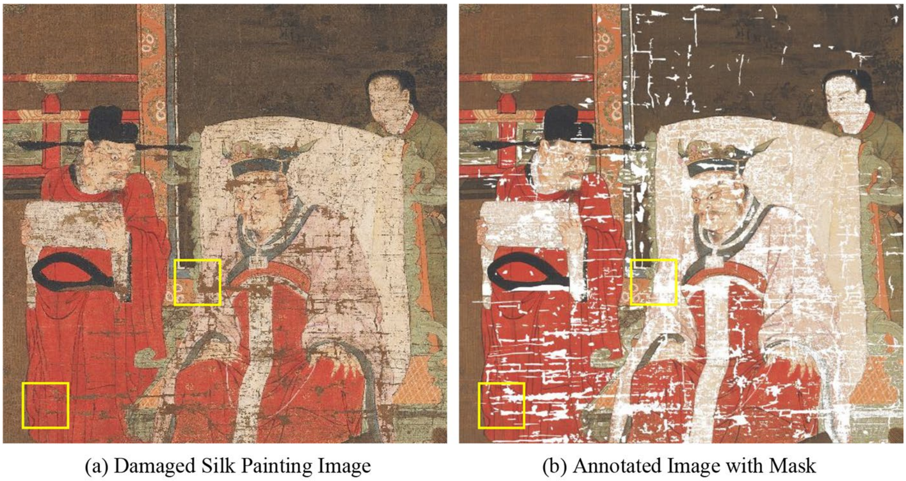


Fig. 6 Qualitative comparisons on simulation random mask (zoom in for a better view)

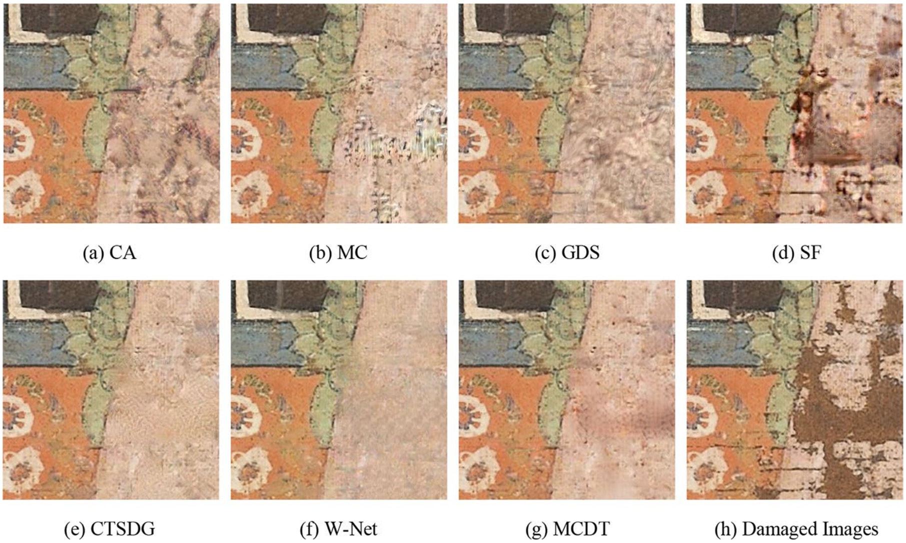
4.6 Application Study

The proposed MCDT model was further extended to silk painting restoration applications in addition to performance validation on Dhmurals1714 dataset. The application study was conducted using authentic damaged silk painting images provided by Gansu Museum and the Automation Research Institute of Gansu Academy of Sciences. The damaged silk painting represents a Sect. (4096 × 4096 pixels) of a Ming Dynasty "Ten Kings" artwork, exhibiting various types of deterioration including punctate, linear, gelatinous, and large-area damage as shown in Fig. 7. The damaged regions were first carefully annotated with masks. And then, both the masked image and original damaged image were partitioned into patches which were processed by different inpainting models. Finally, the restored patches were reassembled to their original dimensions. The proposed MCDT method was compared against six state-of-the-art approaches: CA [4], MC [7], CTSDG [30], SF [31], GDS [32], and W-Net [29]. Magnified views of selected regions (yellow frames in figures) are presented in Figs. 8 and 9 for detailed comparison. CA produces severe blurring artifacts. MC exists over-inpainting with conspicuous black spots. GDS exhibits significant pseudo-artifacts. SF fails to generate plausible content. While CTSDG generates seemingly reasonable results, noticeable ripple artifacts are observed. W-Net results show severe inconsistency with the background as shown in Fig. 9. MCDT produces more natural restoration results. It recovers clearer structures and handles all damage types well, including punctate, linear, gelatinous, and large-area damages. Most importantly, MCDT maintains better visual consistency with the original artwork than other methods.


(a) Damaged Silk Painting Image

(b) Annotated Image with Mask

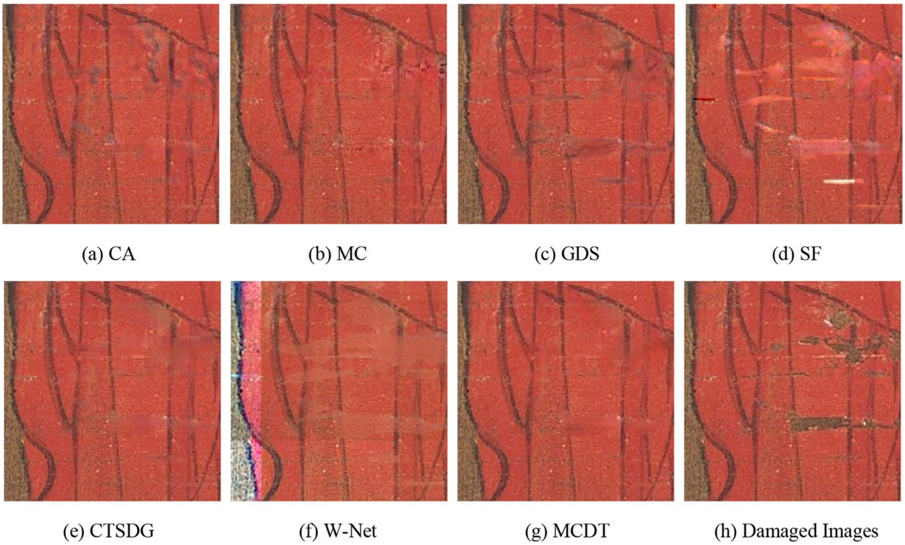
Fig. 7 Damaged silk painting image and annotated mask regions



(a) CA (b) MC (c) GDS (d) SF

(e) CTSDG (f) W-Net (g) MCDT (h) Damaged Images

Fig. 8 Visual comparison between MCDT and other methods on silk paintings (zoom in for a better view)



(a) CA (b) MC (c) GDS (d) SF

(e) CTSDG (f) W-Net (g) MCDT (h) Damaged Images

Fig. 9 Visual comparison between MCDT and other methods on silk paintings (zoom in for a better view)

4.7 User Study

A user study was conducted to evaluate the visual quality of generated results from different approaches. Thirty representative images were selected from the Dhmurals1714 test dataset as evaluation samples. Various benchmark methods were per-

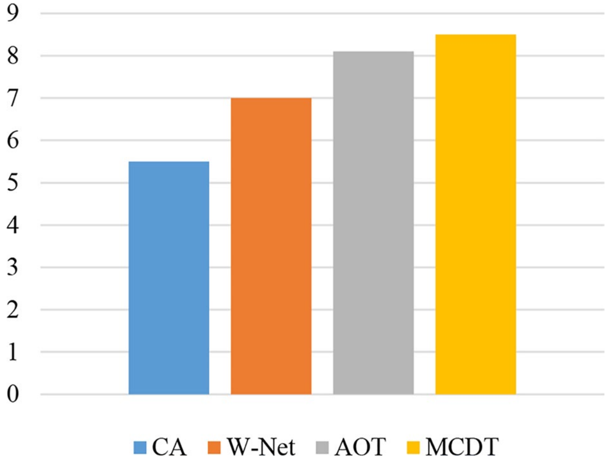


Fig. 10 Statistical results of user study

formed to obtain comparative results for each test image. These restored images were then distributed to 25 research team members for subjective assessment. Participants were instructed to rate the results on a 0–10 scale based on three key quality metrics with higher scores indicating better perceived quality. Quality metrics conclude visual realism, detail richness, and edge sharpness. The average scores were calculated after collecting all evaluations. The reconstruction results obtained by MCDT method were consistently rated higher than those produced by other models as shown in Fig. 10.

5 Conclusion

A Multi-column Condition Decoding Transformer (MCDT) architecture is proposed for cultural relic image inpainting, aiming to incorporate prior knowledge into cultural relic image inpainting. The MCDT framework is designed to integrate multiple prior knowledge, into the Transformer architecture through cross attention mechanisms implemented in multiple decoder branches. The prior knowledge includes detail features and edge features. The multi-column conditional decoder establishes complete mapping relationships between diverse input and output modalities. Verification experiments were carried out on the cultural relic image datasets. The experiment results show that the proposed MCDT model is competent for the cultural relics image inpainting task, and its performance is superior to the existing state-of-the-art methods.

Acknowledgements This work was supported in part by the National Nature Science Foundation of China under Grant 62261032, and the Central Government Guiding Funds for Local Science and Technology Development Program under Grant 25ZYJA026.

Author Contributions C.S. and W.L. wrote the main manuscript text, Z.L. and J.Y. prepared Figs. 7, 8 and 9. All authors reviewed the manuscript.

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

Declarations

Competing Interests The authors declare no competing interests.

References

1. Liu, G., Dundar, A., Shih, K. J., Wang, T. C., Reda, F. A., Sapra, K., Yu, Z., Yang, X., Tao, A., & Catanzaro, B. (2023). Partial convolution for padding, inpainting, and image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(5), 6096–6110. https://doi.org/10.1109/TPAMI.2022.3209702
2. Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In: *Advances in Neural Information Processing Systems*, pp. 2672–2680.
3. Liao, L., Xiao, J., Wang, Z., Lin, C. W., & Satoh, S. (2021). Image inpainting guided by coherence priors of semantics and textures. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6539–6548.
4. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2018). Generative image inpainting with contextual attention. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5505–5514.
5. Iizuka, S., Serra, E. S., & Ishikawa, H. (2017). Globally and locally consistent image completion. *ACM Trans Graph*, *36*(4), 14. https://doi.org/10.1145/3072959.3073659
6. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: feature learning by inpainting. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2536–2544.
7. Wang, Y., Tao, X., Qi, X., Shen, X., & Jia, J. (2018). Image inpainting via generative multi-column convolutional neural networks. In: *Advances in Neural Information Processing Systems*, pp. 331–340.
8. Kim, J., Kim, W., Oh, H., & Lee, S. (2023). Progressive contextual aggregation empowered by Pixel-Wise confidence scoring for image inpainting. *IEEE Transactions on Image Processing*, *32*, 1200–1214. https://doi.org/10.1109/TIP.2023.3238317
9. Nazeri, K., Ng, E., Joseph, T., Qureshi, F. Z., & Ebrahimi, M. (2019). EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning. In: *IEEE International Conference on Computer Vision*, pp. 3265–3274.
10. Zeng, Y., Fu, J., Chao, H., & Guo, B. (2019). Learning pyramid-context encoder network for high-quality image inpainting. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1486–1494 https://doi.org/10.1109/CVPR.2019.00158
11. Shi, C., Liu, W., Meng, J., Jia, X., & Liu, J. (2025). Self-prior guided generative adversarial network for image inpainting. *Visual Computer*, *41*(4), 2939–2951. https://doi.org/10.1007/s00371-024-03578-x
12. Liu, W., Cao, C., Liu, J., Ren, C., Wei, Y., & Guo, H. (2021). Fine-grained image inpainting with scale-enhanced generative adversarial network. *Pattern Recognition Letters*, *143*, 81–87. https://doi.org/10.1016/j.patrec.2020.12.008
13. Zheng, C., Cham, T. J., Cai, J., & Phung, D. (2022). Bridging global context interactions for high-fidelity image completion. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11502–11512.
14. Li, W., Lin, Z., Zhou, K., Qi, L., Wang, Y., & Jia, J. (2022). MAT: Mask-aware transformer for large hole image inpainting. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10748–10758.
15. Wang, Q., Wang, Z., Zhang, X., & Feng, G. (2024). Art image inpainting with Style-Guided Dual-Branch inpainting network. *Ieee Transactions on Multimedia*, *26*, 8026–8037. https://doi.org/10.1109/TMM.2024.3374963
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: transformers for image recognition at scale. In: *International Conference on Learning Representations*, arXiv:2010.11929.

17. Xia, B., Zhang, Y., Wang, S., Wang, Y., Wu, X., Tian, Y., Yang, W., & Gool, L. V. (2023). DiffIR: efficient diffusion model for image restoration, presented at the IEEE International Conference on Computer Vision.
18. Antonio, C., Patrick, P., & Kentaro, T. (2004). Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, *13*(9), 1200–1212.
19. Yu, T., Lin, C., Zhang, S., You, S., Ding, X., Wu, J., & Zhang, J. (2019). End-to-End partial convolutions neural networks for dunhuang grottoes wall-painting restoration, presented at the IEEE international conference on computer vision.
20. Yu, T., Lin, C., Zhang, S., Wang, C., Ding, X., An, H., Liu, X., Qu, T., Wan, L., You, S., Wu, J., & Zhang, J. (2022). Artificial intelligence for Dunhuang cultural heritage protection: The project and the dataset. *Int J Comput Vis*, *130*(11), 2646–2673.
21. Zhou, Z., Liu, X., Shang, J., Huang, J., Li, Z., & Jia, H. (2022). Inpainting digital Dunhuang murals with Structure-Guided deep network. *ACM Journal on Computing and Cultural Heritage*, *15*(4), 1–25. https://doi.org/10.1145/3532867
22. 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/10.1109/TVCG.2022.3156949
23. Liang, J., Cao, J., Sun, G., Zhang, K., Gool, L. V., & Timofte, R. (2021). SwinIR: image restoration using swin transformer. In: *IEEE International Conference on Computer Vision*, pp. 1833–1844 https://doi.org/10.1109/ICCVW54120.2021.00210
24. Dosovitskiy, A., Beyer, L., Kolesnikov, A., DirkWeissenborn, Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In: International *Conference on Learning Representations*.
25. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 15979–15988 https://doi.org/10.1109/CVPR52688.2022.01553
26. Dong, Q., Cao, C., & Fu, Y. (2022). Incremental transformer structure enhanced image inpainting with masking positional encoding. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11358–11368.
27. Zhou, Y., Barnes, C., Shechtman, E., & Amirghodsi, S. (2021). TransFill: reference-guided image inpainting by merging multiple color and spatial transformations. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2266–2276.
28. Li, L., Zou, Q., Zhang, F., Yu, H., Chen, L., Song, C., Huang, X., & Wang, X. (2022). Line drawing guided progressive inpainting of mural damages. arXiv:2211.06649.
29. Zhang, R., Quan, W., Zhang, Y., Wang, J., & Yan, D. M. (2023). W-Net: Structure and texture interaction for image inpainting. *IEEE Transactions on Multimedia*, *25*, 7299–7310. https://doi.org/10.1109/TMM.2022.3219728
30. Guo, X., Yang, H., & Huang, D. (2021). Image inpainting via conditional texture and structure dual generation. In: IEEE International Conference on Computer Vision, 14114–14123.
31. Ren, Y., Yu, X., Zhang, R., Li, T. H., Liu, S., & Li, G. (2019). StructureFlow: image inpainting via structure-aware appearance flow. In: *IEEE International Conference on Computer Vision*, pp. 181–190 https://doi.org/10.1109/ICCV.2019.00027
32. Peng, J., Liu, D., Xu, S., & Li, H. (2021). Generating diverse structure for image inpaintingWith hierarchical VQ-VAE. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10775–10784.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Global Cross Attention Transformer for Image Super-Resolution

Changhong Shi, Weirong Liu(✉), Jiahao Meng, Zhijun Li, and Jie Liu

College of Electric and Informational Engineering, Lanzhou University of Technology,
Lanzhou 730050, Gansu, China
liuwr@lut.edu.cn

Abstract. Transformer-based models have demonstrated state-of-the-art results in the field of image super-resolution. However, we observe that such methods sometimes suffer from overly smooth structural reconstruction and blurred details, indicating that the potential of Transformers has not yet been fully exploited in existing networks. To leverage more prior information, this paper proposes a novel Global Cross Attention Transformer (GCAT) algorithm. This algorithm introduces external prior information by incorporating a cross-attention mechanism alongside the original self-attention mechanism. Furthermore, to better establish the model, we apply cross-attention across all Transformer modules to enhance the model capability for complex mapping. Extensive experiments demonstrate the efficacy of the proposed architecture, with the overall approach exceeding the performance of current state-of-the-art methods.

Keywords: Image super-resolution · Transformer · Cross attention · Prior information

1 Introduction

Single-image super-resolution (SR) is a fundamental problem in image processing and computer vision, focusing on reconstructing high-resolution images from low-resolution inputs. This task is crucial for enhancing image quality in various applications, such as medical imaging, surveillance, and digital photography. Over the years, numerous methods have been developed, ranging from traditional interpolation-based techniques to advanced deep learning models. These approaches aim to recover fine details and improve visual fidelity, making the reconstructed images more useful for downstream tasks. The challenge lies in accurately predicting high-frequency details that are lost during the degradation process, while avoiding artifacts and maintaining natural image appearance [1]. With the successful application of deep learning in SR tasks, methods based on convolutional neural networks (CNNs), such as SRCNN and RCAN, have almost dominated this field in recent years [2–5]. Nevertheless, the impressive achievements of Transformer in natural language processing have increasingly drawn the interest of the computer vision field. As a result, Transformer-based approaches have

L. Jin and L. Wang (Eds.): ISNN 2025, LNCS 15951, pp. 161–171, 2026.
https://doi.org/10.1007/978-981-95-1233-1_15

been adopted for super-resolution tasks [6–12]. For example, SwinIR method has achieved breakthrough improvements in super-resolution tasks [6].

Transformer networks leverage the long-range dependency modeling capability of the self attention to establish effective mappings between inputs and outputs [13]. Although SwinIR achieves superior quantitative performance, its subjective results are sometimes inferior to those of convolutional neural network methods due to its limited scope of information utilization. These observations suggest that Transformers possess a stronger capability for modeling local and long-range information, but their range of information utilization is constrained, requiring additional external mechanisms to expand their scope. However, most current Vision Transformer (ViT) methods only utilize one of the two key modules in Transformers—the self-attention mechanism—for image super-resolution tasks, while neglecting the other key module—the cross-attention mechanism. This lack of external condition integration leads to visually unnatural reconstruction results.

To address the aforementioned issues and explore the potential of leveraging prior information in SR Transformers, we propose a Global Cross-Attention Transformer, namely GCAT. Our GCAT integrates both cross-attention and self-attention mechanisms to harness the ability of cross-attention to incorporate external information and the capability of self-attention for internal representation. Furthermore, to better establish the model, we apply cross-attention across all Transformer modules to enhance the model's capacity for complex mapping.

In summary, our contributions encompass three aspects:

1) We introduce cross-attention into Transformer architecture to utilize more external input information.
2) We apply both cross-attention and self-attention across all Transformer modules to strengthen the model's ability for complex mapping.
3) The proposed GCAT architecture has demonstrated its effective ability to en-hance image reconstruction performance through extensive experiments.

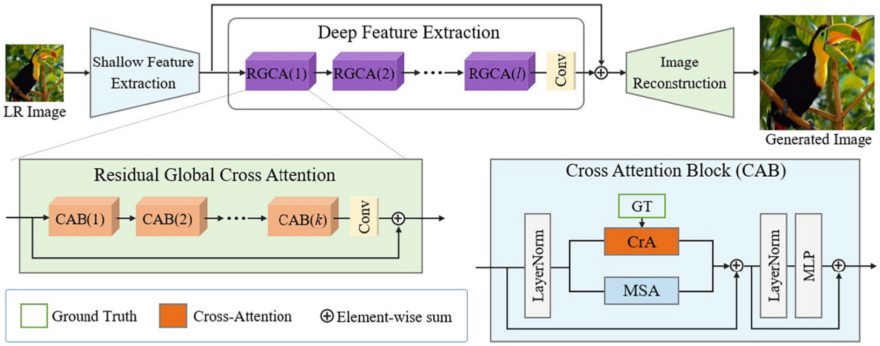


Fig. 1. The overall architecture of GCAT and the structure of CAB.

2 Related Work

2.1 Deep CNN for Super-Resolution

To tackle the issue of image super-resolution (SR), numerous algorithms and mod-els have been proposed. Compared to traditional model-based image restoration methods, CNN-based approaches have become increasingly popular due to their powerful performance. SRCNN was the first to introduce deep convolutional neural networks (CNNs) to the image SR task, achieving superior performance over traditional SR methods. Subsequently, numerous deep learning architectures have been developed for SR to further improve reconstruction quality [5,14]. For instance, more complex convolutional models have been designed and studied to enhance the model's representational capacity, such as residual networks, densely connected networks, and recursive networks. Some works have explored diverse frameworks, including recurrent neural networks, dense link network, and graph neural networks. To improve perceptual quality, adversarial learning has been introduced in to generate more realistic results. By incorporating attention mechanisms, further improvements have been achieved in reconstruction fidelity. Recently, a series of Transformer-based networks have been proposed, continuously pushing the state-of-the-art in SR tasks and demonstrating the powerful long range representational capabilities of Transformers.

2.2 Vision Transformer

Recently, Transformer [15] has demonstrated significant success in natural language processing, leading to the development of various Transformer-based approaches for computer vision tasks, such as image classification, image restoration, object detection, image segmentation, and so on [4,16,17]. Although vision Transformers have demonstrated their superiority in modeling long-range dependencies [18], many studies still suggest that convolutions can help Transformers achieve better visual representation [4]. To address such issues, SwinIR proposed a Transformer-based approach for image restoration. [7] designed a network based on the Vision Transformer (ViT) architecture and applied multi-task pre-training to enhance image processing performance. However, existing methods have yet to fully exploit the potential of Transformers, whereas our approach leverages cross-attention to incorporate more external input information for achieving superior reconstruction results.

3 Method

3.1 Network Architecture

The GCAT framework comprises three components: shallow feature extraction module, deep feature extraction module and image reconstruction module. Figure 1 for an overview. Given a low-resolution input I_{LR}, we use a convolutional layer $H_{SF}(\cdot)$ to extract shallow feature F_0 as:

$$F_0 = H_{SF}(I_{LR}) \tag{1}$$

Obtaining a more stable optimization process is one of the advantages of convolution. It also excels at mapping input images from a low-dimensional feature space to a higher-dimensional feature space. Then, the deep feature F_{DF} is extracted from F_0 as:

$$F_{_{DF}} = H_{_{DF}}(F_{_0}) \tag{2}$$

where $H_{DF}(\cdot)$ is the deep feature extraction operation and it contains l Residual Global Cross Attention (RGCA) and a convolutional layer. Intermediate features and the deep feature output F_{DF} are extracted progressively as follows:

$$\begin{gathered} F_i = H_{\mathrm{RGCA}_i}(F_{i-1}),\ i = 1, 2, \ldots, l \\ F_{\mathrm{DF}} = H_{\mathrm{Conv}}(F_l). \end{gathered} \tag{3}$$

where $H_{RGCA_i}(\cdot)$ denotes the i-th RGCA block and $H_{Conv}(\cdot)$ is the last convolutional layer. The high-resolution result is then reconstructed by image reconstruction module.

3.2 Residual Global Cross Attention (RGCA)

As shown in Fig. 1, each RGCA contains k cross attention blocks (CAB) and a 3×3 convolutional layer. Intermediate features $F_{i,1}$, $F_{i,2}$, ... , $F_{i,k}$ by k cross attention blocks are first extracted. The i-th RGCA can be formulated as:

$$\begin{gathered} F_{i,j} = H_{CAB_{i,j}}(F_{i,j-1}), j = 1, 2, ...k \\ F_{i,out} = H_{Con\nu_i}(F_{i,k}) + F_{i,0} \end{gathered} \tag{4}$$

where $H_{CAB_{i,j}}(\cdot)$ is the j-th cross attention block in the i-th RGCA.

3.3 Cross Attention Block(CAB)

The cross-attention mechanism facilitates cross-modal interaction among different inputs, where the primary mapping relationship we aim to establish is between the low-resolution image and its high-resolution ground truth counterpart. Therefore, we compute cross-attention between the ground truth and low-resolution features and integrate the results with the standard Transformer block, as illustrated in Fig. 2. Specifically, the CrA layer's input consists of three components: one from the ground truth and the other two from the previous hidden layer. These inputs undergo convolution and linear operations to generate three feature spaces as follows:

$$\begin{gathered} f = \boldsymbol{W}_f H_{Con\nu}(F_{gt_{i,out}}) \\ g = \boldsymbol{W}_g \cdot H_{Conv}(F_{i,j}) \\ h = \boldsymbol{W}_h \cdot H_{Con\nu}(F_{i,j}) \end{gathered} \tag{5}$$

The cross attention weights reflect how much the model focuses on individual pixels of the ground truth while generating corresponding pixels in the hidden features. These weights are derived through the following calculation:

$$s_{cr} = Soft\max(fg^{\mathrm{T}}) \tag{6}$$

The output of cross attention layer $o = W_O s_{sr} h$, where W_f, W_g, W_h, and W_o are learnable weight matrices.

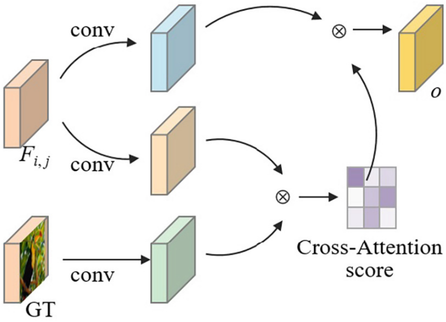


Fig. 2. The structure of cross-attention (CrA) module.

4 Experiments

4.1 Experimental Setup

During both training and testing, the proposed method GCAT in this paper utilizes a hardware platform consisting of an Intel(R) Core(TM) i7-8700 CPU (3.2 GHz) and a single NVIDIA GeForce RTX 4090 GPU (24 GB). The environment is: Windows 10, Python 3.7, PyTorch 1.7.1, CUDA v11.0, and cuDNN v11.0. We use DIV2K dataset as the original training dataset. The number of RGCA and CAB blocks are both set to 6 for the structure of GCAT. The attention head number are set to 6 for CrA and MSA. The metrics PSNR and SSIM are reported to evaluate the quantitative performance. Note that, in the testing phase, the input to the cross attention layers is derived solely from the hidden layer features and does not include the ground truth image.

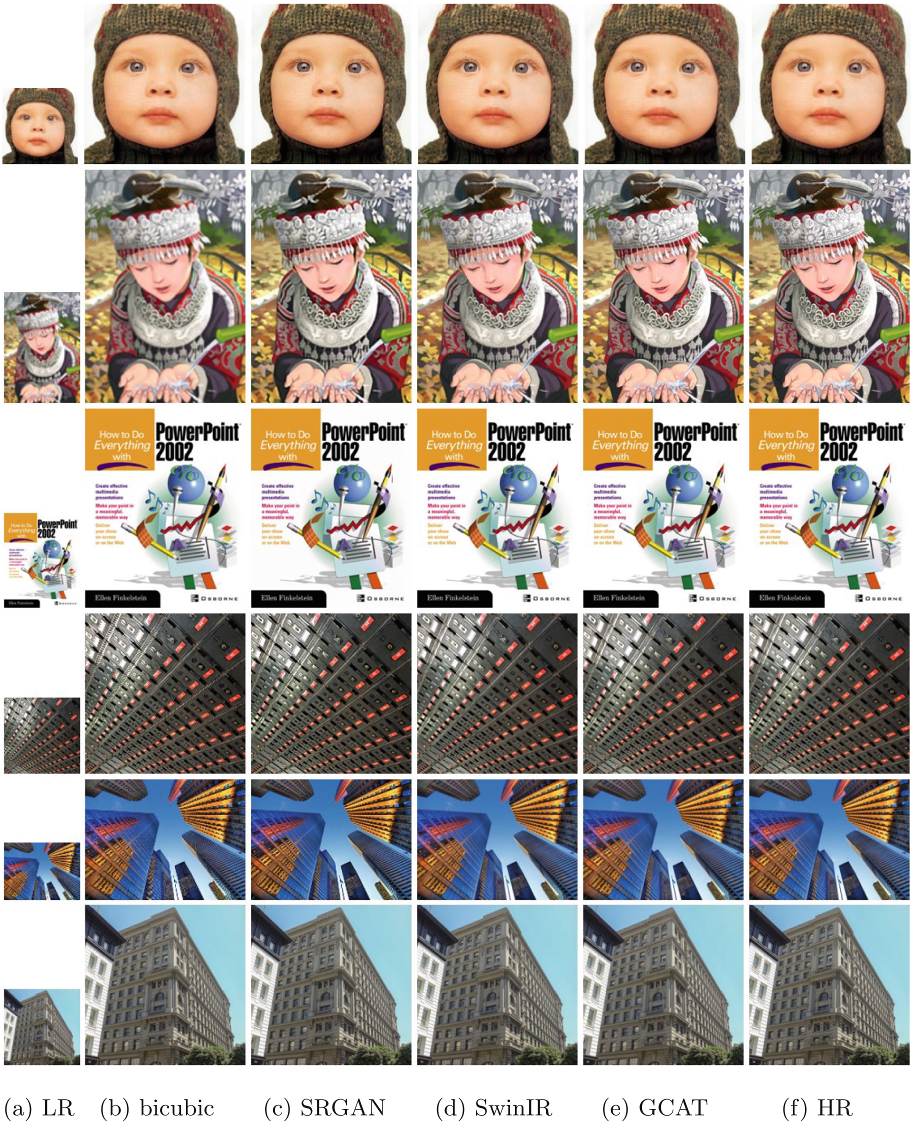
4.2 Comparison with State-of-the-Art Methods

Quantitative Results. We conducted quantitative comparison experiments between the proposed algorithm, GCAT, and two representative methods, SRGAN and SwinIR, on four different datasets: Set5, Set14, BSD100, and Urban100, using scaling factors of 2× and 4×. Table 1 presents the quantitative comparison results of our method against the others. As shown in Table 1, our method outperforms the others on almost all benchmark datasets, with superior average metrics. Particularly, for 2× super-resolution, the PSNR value of GCAT on the Set14 dataset is 0.69 dB higher than that of SwinIR. More notably, for 4× super-resolution, the SSIM value of GCAT on the Urban100 dataset is 24% higher than that of SwinIR, and the PSNR value is 32% higher.

Table 1. Quantitative comparison with state-of-the-art methods for image SR on four benchmark datasets.

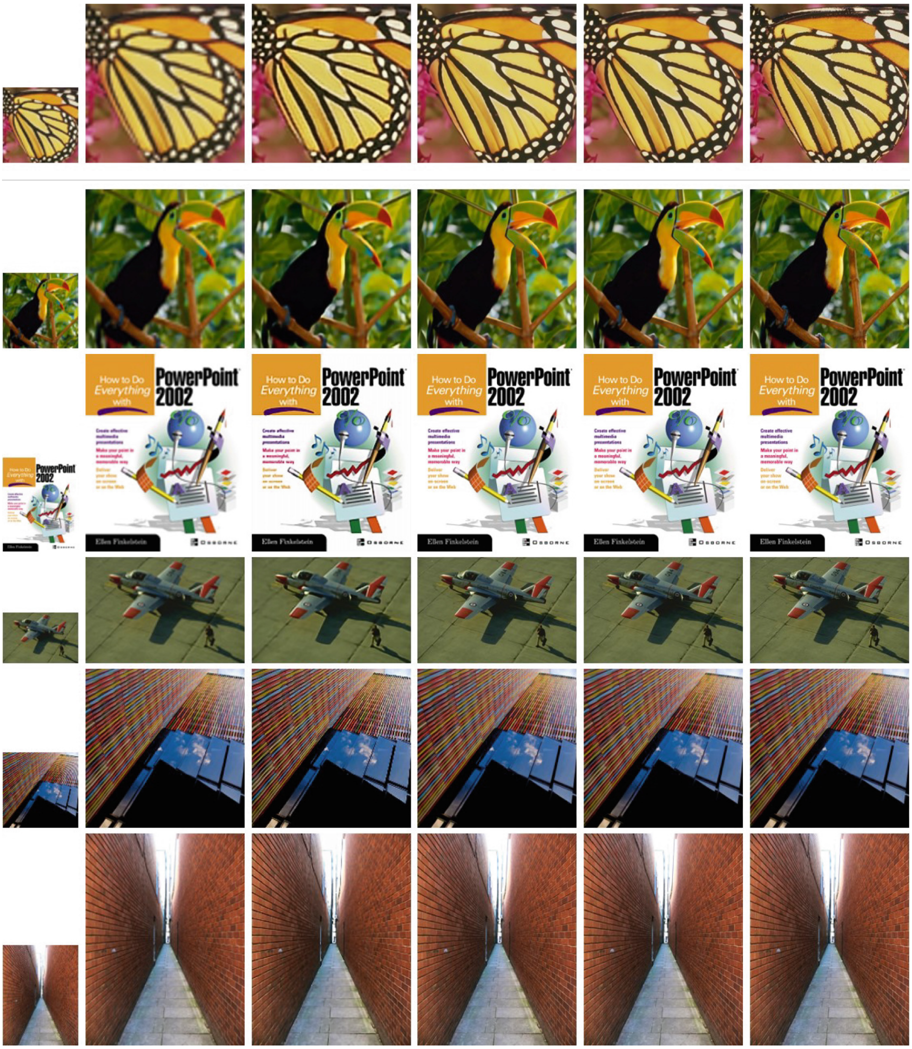
Methods	Scale	Set5		Set14		BSD100		Urban100	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	×2	33.66	0.9299	30.24	0.8688	29.56	0.8431	26.88	0.8403
SRGAN	×2	33.22	0.9301	30.66	0.8855	31.76	0.8931	31.11	0.9173
SwinIR	×2	38.35	0.9620	34.14	0.9227	32.44	0.9030	33.40	0.9393
GCAT (ours)	×2	**38.60**	**0.9629**	**34.83**	**0.9265**	**32.61**	**0.9050**	**33.62**	**0.9424**
Bicubic	×4	24.40	0.6580	23.10	0.5660	23.67	0.5480	20.74	0.5160
SRGAN	×4	29.98	0.8411	26.68	0.7178	24.50	0.7364	25.74	0.6668
SwinIR	×4	32.72	0.9021	28.94	0.7914	27.83	0.7459	27.07	0.8164
GCAT (ours)	×4	**33.04**	**0.9055**	**29.21**	**0.7971**	**27.99**	**0.7515**	**27.94**	**0.8358**

Visual Comparison. We conducted subjective comparison experiments with different magnification factors on four distinct datasets: Set5, Set14, BSD100, and Urban100. The selected test images include 'baby', 'bird', and 'butterfly' from Set5, 'baboon', 'comic', 'PPT', and 'zebra' from Set14, 'BSD100_071' from BSD100 dataset, as well as 'img_006,' 'img_012,' and 'img_014', 'img_023', and 'img_038' from Urban100 dataset. The results are shown in Fig. 3 and Fig. 4, where Fig. 3 corresponds to a magnification factor of 2, and Fig. 4 corresponds to a magnification factor of 4. The images from left to right displayed are: the LR image, the bicubic super-resolution result, the result from the SRGAN method, the result from the SwinIR method, our GCAT result, and HR image. It can be observed that our method, GCAT, successfully reconstructs fine-grained details. In contrast, other methods exhibit varying degrees of blurring effects. Specifically, as shown in the second row of Fig. 3, the red box in the ethnic costume image highlights the hair strands near the ear. Compared to the bicubic interpolation method, which produces a very blurry hair structure, the SRGAN method reconstructs a more complete structure but with less sharp edges and noticeable artifacts. On the other hand, our proposed GCAT method achieves clearer edges and richer details in the reconstruction. The GCAT method successfully reconstructs clear grid patterns for the second-to-last row image 'Urban100_012' in Fig. 3, while the other three methods all suffer from blurry effects. Similarly, under high scaling factors, the subjective effects reconstructed by our GCAT still exhibit richer details and clearer structures compared to other methods as shown in the 'PPT' and 'img_38' images of Fig. 4. Therefore, combined with quantitative comparison results, the subjective results further demonstrate the superiority of our method.



(a) LR (b) bicubic (c) SRGAN (d) SwinIR (e) GCAT (f) HR

Fig. 3. Visual comparison for ×2 SR on Set5, Set14, BSD100, and Urban100 datasets.



(a) LR (b) bicubic (c) SRGAN (d) SwinIR (e) GCAT (f) HR

Fig. 4. Visual comparison for ×4 SR on Set5, Set14, BSD100, and Urban100 datasets.

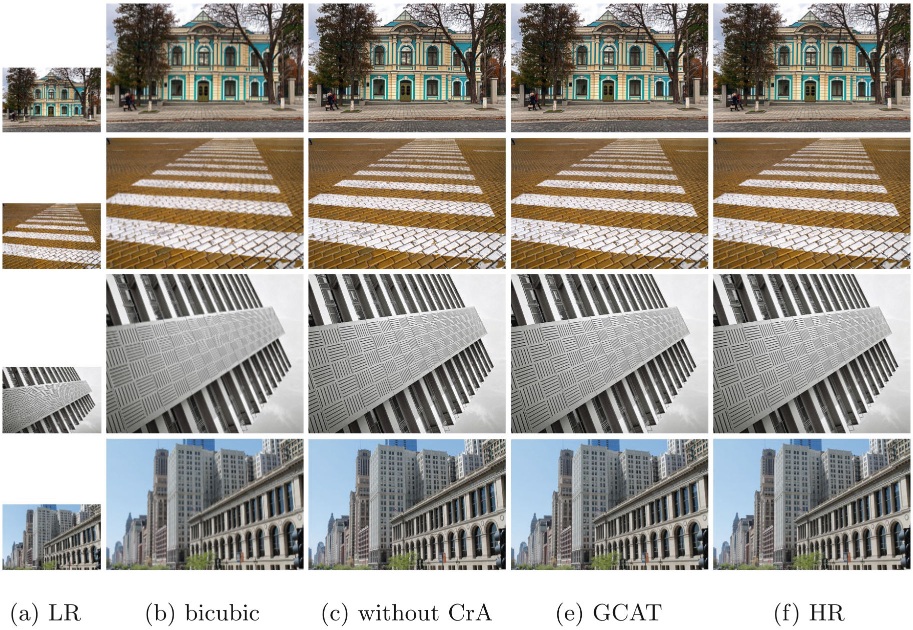


Fig. 5. Visual comparison for ×4 SR on Set5, Set14, BSD100, and Urban100 datasets.

4.3 Ablation Study

This section investigates the contributions of key modules in the GCAT framework through ablation experiments, with a primary focus on validating the effectiveness of the cross-attention mechanism. The model was trained on DIV2K dataset, and testing was conducted on Urban100 dataset. First, a simplified version of the model (with all cross-attention modules removed) was trained. Subsequently, its performance was qualitatively compared with the full GCAT model on the ×3 super-resolution task. The performance differences between the architectures are evident as is shown in Fig. 5. Figure 5(a) shows the input low-resolution image, Fig. 5(b) is the Bicubic interpolation results, which exhibit noticeable edge blurring and detail smoothing. Although the simplified model in Fig. 5(c) shows some improvement over Bicubic, it still suffers from structural distortions and detail blurring in complex texture regions (*e.g.*, the brick patterns on building facades). In contrast, the reconstruction results of the GCAT model in Fig. 5(d) not only maintain more accurate geometric structures but also significantly outperform other methods in terms of detail richness and edge sharpness. Particularly, the reconstruction of regular repetitive texture patterns closely approximates the high-resolution image in Fig. 5(e). By establishing global dependencies between Transformer submodules and external priors, the model can more accurately reconstruct high-frequency details and sharp edges.

5 Conclusion

In this paper, we address the limitations of existing Transformer-based methods, which often exhibit overly smooth structural reconstruction and blurred details, suggesting that the full potential of Transformers has not been fully harnessed. To leverage more prior information, we propose a novel Global Cross Attention Transformer (GCAT) algorithm. The GCAT model enhances the traditional self-attention mechanism by integrating a cross-attention mechanism, thereby introducing external prior information. Additionally, cross-attention is applied across all Transformer modules to strengthen the complex mapping capability. Extensive experimental results validate the effectiveness of the proposed module, demonstrating that our overall method surpasses state-of-the-art approaches in performance.

References

1. Anwar, S., Khan, S., Barnes, N.: A deep journey into super-resolution: a survey. ACM Comput. Surv. **53**(3), 1–21 (2020)
2. Zhang, Y., Li, K., Li, K.,Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: 2018 European Conference on Computer Vision, pp. 294–310. Springer, Cham (2018)
3. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 184–199. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_13
4. Lepcha, D.C., Goyal, B., Dogra, A., Goyal, V.: Image super-resolution: a comprehensive review, recent trends. Challenges Appl. Inf. Fusion. **91**, 230–260 (2023)
5. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2015)
6. Liang, J., Cao, J., Sun, G., Zhang, K., Gool, L.V., Timofte, R.: SwinIR: image restoration using swin transformer. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops, pp. 1833–1844. IEEE (2021)
7. Chen, H., et al.: Pre-trained image processing transformer. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12294–12305. IEEE (2021)
8. Chen, X., Wang, X., Zhou, J., Dong, C.: Activating more pixels in image super-resolution transformer. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22367–22377. IEEE (2023)
9. Lei, S., Shi, Z., Mo, W.: Transformer-based multistage enhancement for remote sensing image super-resolution. IEEE Trans. Geosci. Remote Sens. **60** (2022)
10. Liang, Z., Wang, Y., Wang, L., Yang, J., Zhou, S.: Light field image super-resolution with transformers. IEEE Signal Process. Lett. **29**, 563–567 (2022)
11. Lu, Z., Li, J., Liu, H., Huang, C., Zhang, L., Zeng, T.: Transformer for single image super-resolution. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 456–465 (2022)
12. Chen, Z., Zhang, Y., Gu, J., Kong, L., Yang, X., Yu, F.: Dual aggregation transformer for image super-resolution. In: IEEE International Conference on Computer Vision, pp. 12278–12287. IEEE (2023)

13. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: a survey. ACM Comput. Surv. **54**(10s), 1–41 (2022)
14. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 391–407. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_25
15. Vaswani, A., et al.: Attention is all you need. In: 31st International Conference on Neural Information Processing Systems, pp. 6000–6010. Curran Associates Inc. (2017)
16. Liu, A., Liu, Y., Gu, J., Qiao, Y., Dong, C.: Blind image super-resolution: a survey and beyond. IEEE Trans. Pattern Anal. Mach. Intell. **45**(5), 5461–5480 (2023)
17. Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22367–22377. IEEE (2023)
18. Parmar, N., et al.: Image Transformer. In: 35th International Conference on Machine Learning, pp. 4055–4064. PMLR (2018)