兰州理工大学

科研成果汇总

学　　号：	221081104001
研 究 生：	孟家豪
导　　师：	刘微容 教授
研究方向：	图像处理与模式识别
论文题目：	基于感受野和先验信息增强的图像修复方法及其应用研究
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：	2022 年 9 月

2025 年 1 1 月 1 日

目　录

1. 论文检索报告 1
2. 论文录用证明 4
3. Meng Jiahao, Liu Weirong, Shi Changhong, Li Zhijun, Liu Chaorong. Degression receptive field network for image inpainting[J]. Engineering Applications of Artificial Intelligence, 2024, 138:109397.(SCI: WOS:001339192900001) 5
4. Meng Jiahao, Liu Weirong, Shi Changhong, Li Zhijun, Liu Jie. Self-information and prediction mask enhanced blind inpainting network for dunhuang mural images[J]. Engineering Applications of Artificial Intelligence, 2025, 159:111769. (SCI: WOS:001554583100001).....22
5. Meng Jiahao, Liu Weirong, Shi Changhong, Li Zhijun, Liu Jie. Multi-receptive fields and dynamic matching of damaged patterns for image inpainting[J/OL]. Journal of Southeast University(English Edition) , 2025: 1-22.(EI 源刊，已录用) 35

图书馆 文献检索报告

兰州理工大学图书馆

报告编号：R2025-1215

机构：兰州理工大学 电气工程与信息工程学院

姓名：孟家豪 [221081104001]

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-2025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	2024 - 2025	2	2
EI 收录	EI-Compendex	2024 - 2025	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）：孟家豪

完 成 人（签字）：徐春园

完 成 日 期 ：2025年10月27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）

兰州理工大学图书馆 信息咨询与学科服务部

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：R2025-1215 SCI-E 收录

数据库：科学引文索引 (Science Citation Index Expanded) 时间范围：2024年至2025年	作者姓名：孟家豪 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：徐春园 检索日期：2025年10月27日

检索结果：被 SCI-E 收录文献 2 篇

#	作者	标题	来源出版物	文献类型	入藏号
1	Meng, JH; Liu, WR; Shi, CH; Li, ZJ; Liu, J	Self-information and prediction mask enhanced blind inpainting network for dunhuang murals	*ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE* 2025, 159: 111769.	J Article	WOS:001545831000001
2	Meng, JH; Liu, WR; Shi, CH; Li, ZJ; Liu, CR	Degression receptive field network for image inpainting	*ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE* 2024, 138: 131158.	J Article	WOS:001391929000001
				合计	2

图书馆

文献检索报告
EI 收录

SEARCH

兰州理工大学图书馆

报告编号：R2025-1215 EI 收录

数据库：工程索引 (Engineering Index) 时间范围：2024年至2025年	作者姓名：孟家豪 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：徐春园 检索日期：2025年10月27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	Meng, Jiahao; Liu, Weirong; Shi, Changhong; Li, Zhijun; Liu, Chaorong	College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou	Degression receptive field network for image inpainting	*Engineering Applications of Artificial Intelligence* 2024, 138: 109397.	Journal article (JA)	20244117173340
2	Meng, Jiahao; Liu, Weirong; Shi, Changhong; Li, Zhijun; Liu, Jie	College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou	Self-information and prediction mask enhanced blind inpainting network for dunhuang murals	*Engineering Applications of Artificial Intelligence* 2025, 159: 111769.	Journal article (JA)	20252918809068
					合计	2

证书编号：2025101422773

中国知网学术期刊网络首发论文出版证书

《中国学术期刊（网络版）》是中国核工业集团有限公司主管、中核战略规划研究总院有限公司主办、国家新闻出版署 2015 年 8 月 14 日批准的网络版连续出版物（国际标准连续出版物号　ISSN 2096-4188；国内统一连续出版物号 CN 11-6037/Z），在《中国学术期刊（光盘版）》电子杂志社有限公司的互联网出版网站中国知网（www.cnki.net，网出证（京）字第 416 号）上进行出版。

《Journal of Southeast University(English Edition)》与《中国学术期刊（光盘版）》电子杂志社有限公司合作，在中国知网创办了与《Journal of Southeast University(English Edition)》内容一致的网络版，以单篇或整期出版形式，在印刷版出版之前刊发纸质期刊已正式录用定稿的论文，著作权受法律保护。论文发表时间按中国知网的网络出版时间确认。

兹有，孟家豪，刘微容，史长宏，李治俊，刘婕同志的题为《基于多感受野与破损模式动态匹配的图像修复方法(英文)》的论文

链接地址：https://link.cnki.net/urlid/32.1325.N.20251014.1441.002

已于 2025 年 10 月 14 日在中国知网出版，出版证书验证地址 www.cnki.net，特此证明。

《Journal of Southeast University(English Edition)》编辑部（签章）

《中国学术期刊（光盘版）》电子杂志社有限公司（签章）

2025 年 10 月 14 日

Engineering Applications of Artificial Intelligence 138 (2024) 109397

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Degression receptive field network for image inpainting

Jiahao Meng, Weirong Liu *, Changhong Shi, Zhijun Li, Chaorong Liu

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

ARTICLE INFO

Keywords:
Image inpainting
Generative adversarial networks
Degression receptive field
Coarse to fine inpainting network
Object removal and image editing
Deep learning

ABSTRACT

—Multi-stage image inpainting methods from coarse-to-fine have achieved satisfactory inpainting results in recent years. However, an in-depth analysis of multi-stage inpainting networks reveals that simply increasing complexity of refined network may lead to degradation problems. The paper proposes a degression receptive field network (DRFNet) via multi-head attention mechanism and U-shaped network with different receptive fields to address above phenomenon that existing image inpainting methods have detail blur and artifacts due to insufficient constraints. Initially, DRFNet innovatively takes receptive field as a perspective and consists of five sub-networks with decreasing receptive fields. Secondly, an easy-to-use TransConv module is designed to overcome the problem of local-pixel influence in convolution. Experiments show that comprehensive optimal rate of DRFNet on L1 error, Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Fréchet Inception Distance (FID), and Learned Perceptual Image Patch Similarity (LPIPS) is more than 82.86% on all three benchmark datasets, which achieves state-of-the-art results. Moreover, real-world experiments demonstrate the potential of DRFNet for object removal and image editing. The code is available at: https://github.com/IPCSRG/DRFNet-Inpainting.git.

1. Introduction

Image inpainting is a technique that extracts information from non-missing regions of a damaged image to infer and fill in missing areas. Since introduction by (Bertalmio et al., 2000), image inpainting has received growing research attention and has become one of fundamental tasks in computer vision, which can provide help for computer vision downstream tasks (Chen et al., 2024b; Gao et al., 2024). Currently, image inpainting is widely used in various fields such as old photo inpainting (Cai et al., 2024), image editing (Dere et al., 2024), and has been extended to video image inpainting (Li et al., 2023), cultural relics inpainting (Zhang et al., 2024).

In early stages, image inpainting was mainly dominated by diffusion-based methods and patch-matching-based methods (Zhang et al., 2023b). Diffusion-based methods (Bertalmio et al., 2003; Shen and Chan, 2002) utilize partial differential equations to propagate gradient information from non-missing areas orthogonally into damaged regions. Patch-matching-based methods (Barnes et al., 2009) search for patches within undamaged areas similar to damaged areas, which are then copied and pasted into damaged areas for filling.

Although these methods achieved particular successes at the time, they still had significant shortcomings. As shown in Fig. 1, the reconstruction of pixels by diffusion-based method is limited to locally available information and cannot recover a reasonable structure in missing regions. Patch-matching-based methods assume structures and textures similar to repaired area can be found elsewhere in image. As a result, these methods excel at repairing backgrounds and tasks involving repetitive structures but struggle with reconstructing non-patterned images.

In recent years, researchers have introduced convolutional neural networks (CNN) (Hinton and Salakhutdinov, 2006) and generative adversarial networks (GAN) (Goodfellow et al., 2014) into image inpainting, which largely solve problems of diffusion-based and patch-matching-based methods. Typically, Pathak (Pathak et al., 2016) first introduced generative adversarial loss to image inpainting and named it context-encoder (CE) network. At that time, CE made a breakthrough in repairing center-square area damage by taking advantage of CNNs and GANs. However, the discriminator's focus on fixed square-shaped losses limits CE to repairing missing areas of a specific size. To address above issues, Yu(Yu et al., 2018) proposed contextual attention (CA) mechanism and further designed a two-stage coarse-to--fine image inpainting network. This network ushered deep learning-based image inpainting methods into a new phase of irregular hole inpainting. However, CA exhibits significant repair traces, such as

* Corresponding author.
E-mail address: liuwr@lut.edu.cn (W. Liu).

https://doi.org/10.1016/j.engappai.2024.109397
Received 2 January 2024; Received in revised form 3 September 2024; Accepted 25 September 2024
Available online 11 October 2024

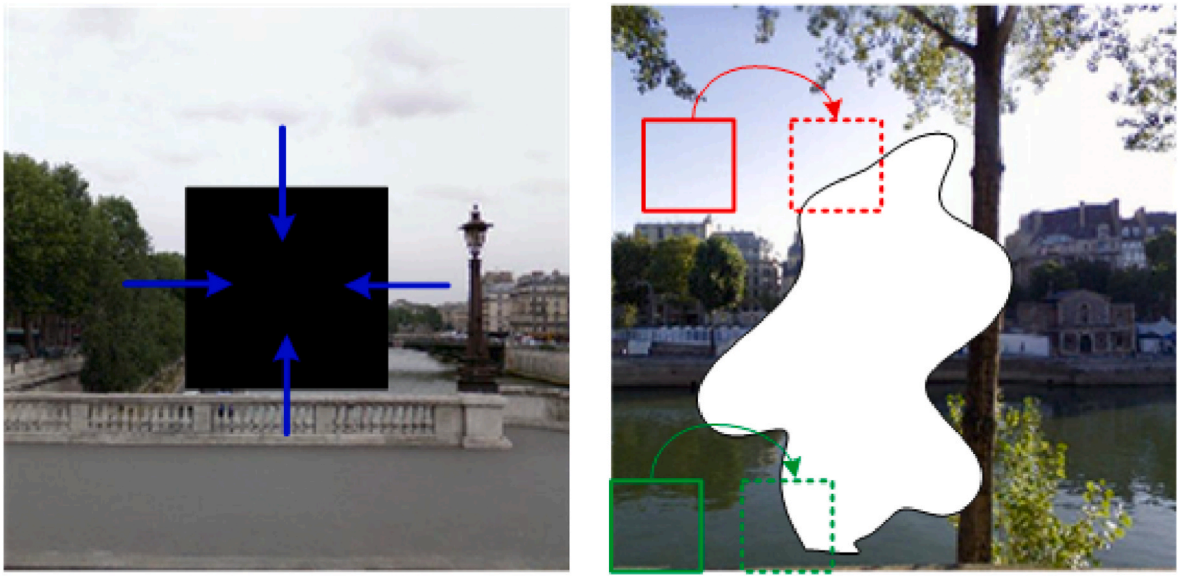


Fig. 1. Diagram of diffusion-based and patch-matching-based methods. The left shows diffusion-based methods, while the right shows patch-matching-based methods.

semantic confusion and structural inconsistency, when applied to large missing regions. Nazeri (Nazeri et al., 2019) divided image inpainting task into two parts: structure prediction and image completion. They developed a novel two-stage image inpainting model, edgeconnect (EC), constrained by edge structure information. The model first generates edge-repaired image and then uses edge image as a constraint to guide image inpainting stage. The introduction of edge structure information enabled EC to achieve a breakthrough in repairing large missing regions with simple structures. However, edge structure extraction is challenging in complex scenes, which limits the application of edge-constrained image inpainting methods. Xu (Xu et al., 2020) proposes a multi-granularity generative adversarial network (MGAN) image inpainting model that improves texture structure and visual continuity through a multi-granularity generative and discriminative strategy. This model adopts a two-stage coarse-to-fine approach, where the reconstructive network initially creates a rough outline, then generative network adds fine-grained details. While MGAN has achieved significant improvements in perceptual-distortion plane for image inpainting, there are still issues with structural disorder in inpainted results, and its capability to handle large irregular masks still needs to be verified. Liu (Liu et al., 2020) addressed the issue of structure and texture inconsistency in current coarse-to-fine inpainting networks and proposed mutual encoder-decoder with feature equalizations (MEDFE) that fuses deep and shallow features. Specifically, MEDFE recovers structural information from deep features and texture information from shallow, then utilizes feature equalization to complement decoder features. Although mutual encoder-decoder improves inpainting quality, repaired areas still show blurry artifacts. To address current two-stage inpainting methods that are prone to fuzzy boundaries and degraded structures as well as underutilization of masks in encoding phase, Zhu (Zhu et al., 2021) presented mask-aware dynamic filtering module (MADF) and point-line normalization, and finally constructed a coarse-to-fine cascaded inpainting model. Experimental analysis reveals that MADF is susceptible to artifacts caused by masks when repairing areas of contiguous damage.

Over past three years, image inpainting methods have grown rapidly due to the advancements in hardware computational capabilities and introduction of various frameworks. Typically, Zheng (Zheng et al., 2022) uses the long-range dependency capabilities of transformer to establish semantic similarity between pixel blocks within image to be repaired and then proposed a novel two-stage coarse-to-fine inpainting network, transformer-based architecture to fill reasonable content (TFill), achieving a breakthrough in semantic consistency for image restoration. However, TFill is susceptible to local mask effects since transformer operates on patches. Quan et al. (2022) proposed a novel three-stage inpainting framework with local and global refinement (LGNet) based on different receptive fields. LGNet first performs coarse inpainting by a codec with large receptive field, then applies a small receptive field shallow network for local refinement, and finally performs attention-based codec for global refinement. LGNet has achieved breakthrough results in image inpainting due to its innovative network architecture. However, its repair results are still plagued by artifacts. Phutke (Phutke and Murala, 2023) proposed a two-stage cascaded inpainting network that uses a spatial projection layer (SPL) in second stage to refine and enhance inpainted results. Specifically, SPL does not rely on attentional mechanism to project spatial contextual information between non-missing and missing regions of image to generate a spatially coherent inpainted image. Despite significant improvements in semantic consistency, the network still produces blur and artifacts when texture information is missing over a large area. Chen (Chen et al., 2023a) proposed a novel multi-scale patch-gan architecture with edge detection (MPGE) for image inpainting, which introduces local and global discriminators to capture high and low-frequency features. MPGE has achieved good repair results for central area repair, but its ability to repair irregular masks has not been validated. Chen (Chen et al., 2023b) proposed a residual feature attention network (RFA-Net), introducing a new backbone architecture for image inpainting. RFA-Net used a residual CNN structure without pooling layers, thus preserving high-frequency details essential for fine-grained texture inpainting. Despite the outstanding effect of RFA-Net in eliminating artifacts in repaired images, there is still a problem with blurred details. Zhang (Zhang et al., 2023a) proposed a w-shaped network (W-Net) with coarse-to-fine image inpainting process. W-Net improves quality of repair results through texture spatial attention (TSA) and structural channel excitation (SCE) modules. The TSA module fills in incomplete textures using reliable attention scores guided by coarse structures, reducing inconsistencies from appearance to semantics, while SCE module corrects structures based on differences between coarse and fine structures, prompting them to produce more reasonable shapes. However, W-Net has problems with inability to maintain semantic consistency when there is insufficient structural information and the possibility of artifacts in repair results. Chen (Chen et al., 2024a)

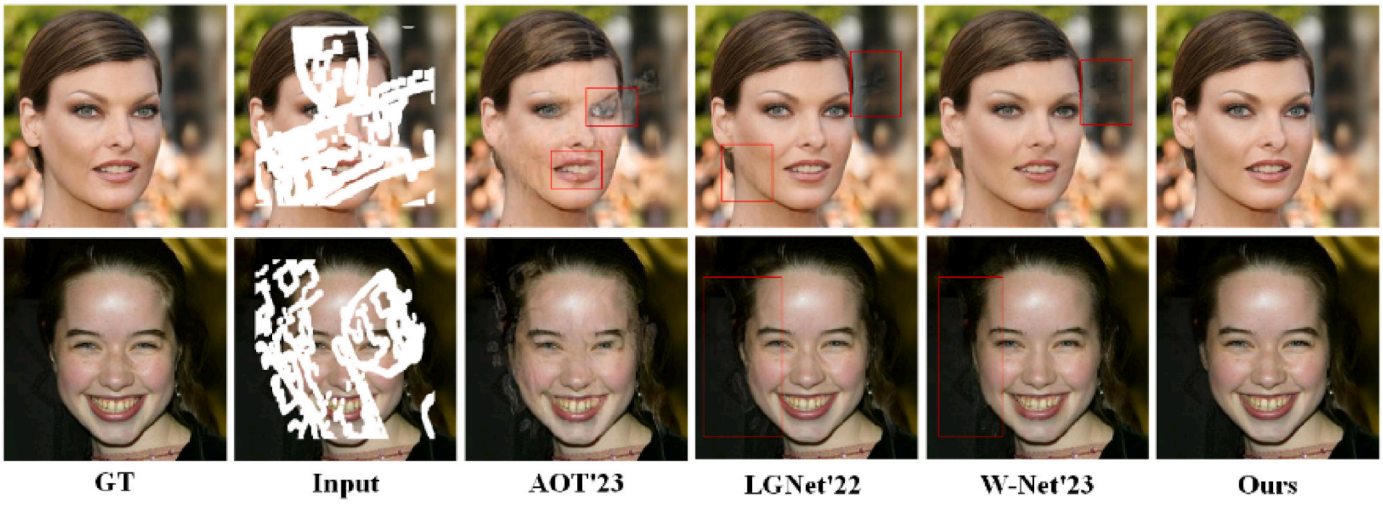


Fig. 2. Examples of single-stage and multi-stage networks. Where AOT is single-stage network, LGNet and W-Net are multi-stage networks. GT represents Ground Truth. Zoom in for more details.

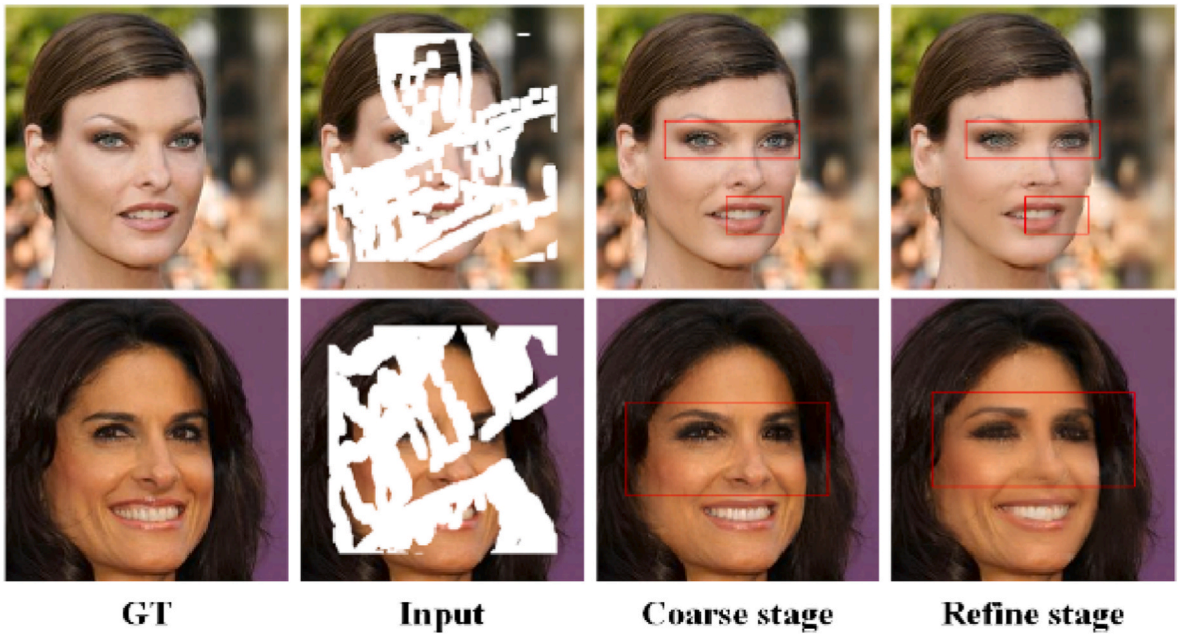


Fig. 3. Analysis of TFill experimental results. Coarse stage is a codec with transformer. GT represents Ground Truth. Zoom in for more details.

proposed an image inpainting model based on partial multi-scale channel attention and residual networks (DNNAM), which can be applied to various scene image inpainting tasks. DNNAM addresses the shortcomings of current deep image inpainting methods in perceiving and expressing of image features at multi scales.

To summarize, image inpainting networks based on deep learning are mainly divided into single-stage and coarse-to-fine multi-stage networks according to design form. Single-stage models typically apply a single codec as generator but still have problems with detail blur, semantic errors, and more severe artifacts due to insufficient constraints. As shown in Fig. 2, single-stage network fails with semantic errors and blurred details when critical information is occluded, such as the eyes and teeth of characters in the third column of first row. Coarse-to-fine multi-stage network typically uses two codecs as generators, with the first codec responsible for coarse repair and the second for detail filling. Multi-stage inpainting networks have stronger constraint capabilities compared to single-stage networks. Fig. 2 shows that semantically reasonable images can be repaired even when critical information is obscured. However, there are still problems with artifacts.

Indeed, the current coarse-to-fine inpainting networks outperform single-stage. Therefore, to better understand the mechanisms of coarse-to-fine inpainting networks, an in-depth analysis was conducted on two representative models, TFill and LGNet. During the analysis of TFill, it was found that there might be a degradation problem in refine stage in coarse-to-fine inpainting architecture, as shown in Fig. 3. The emergence of degradation problems implies that simply increasing the number of network stages does not necessarily lead to better inpainting results. Meanwhile, G2 with a small receptive field in LGNet can effectively refine the output results of G1, while G3 with the same receptive field as G1 has a limited enhancement of repair details, as shown in Fig. 4. In contrast, both G1 and G2 use structure composed of convolutional layers with skip connections, without complex modules. This suggests that network structure with decreases receptive field is feasible. Inspired by above analysis, an image inpainting network with decreasing receptive field (DRFNet) is designed. DRFNet performs multi-scale decreasing receptive fields as network design idea and uses a U-

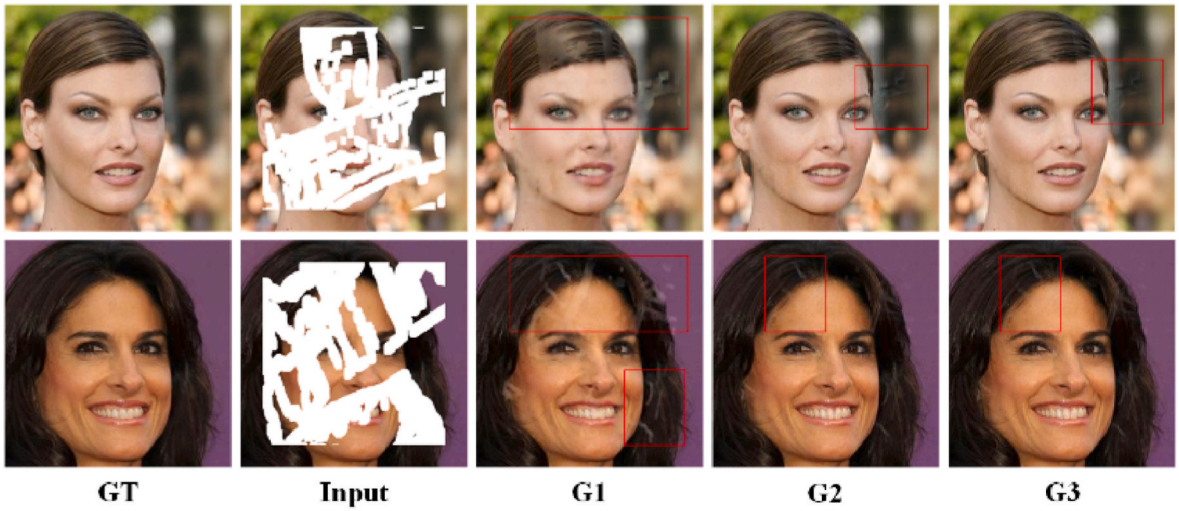


Fig. 4. Analysis of LGNet experimental results. GT represents Ground Truth, G1 is coarse inpainting network (receptive field is 766X766), G2 is local refinement network (receptive field is 109x109), and G3 is global refinement network (receptive field is 766x766). Zoom in for more details.

shaped network with large receptive field as first codec. Then, gradually reduce number of layers of U-shaped network to reduce receptive field. In addition, due to pure convolutional networks have the problem of local influence of pixels, which can be solved to a certain extent by long-range dependency capability of transformer. Therefore, to address the problem of local influence, TransConv module has been designed by integrating a redesigned and simplified transformer with convolutional operations. As shown in Fig. 2, DRFNet has better artifact removal and inpainting abilities that benefit from degression receptive field and TransConv.

Aiming at the phenomenon of detail blur and artifacts in the current multi-stage image inpainting model. An in-depth analysis of representative multi-stage networks is performed and found that simply increasing the number and complexity of sub-networks may lead to degradation problem. To address above problem, a degression receptive field network is proposed, which is a new design idea for image inpainting tasks. For this paper, the main contributions are as follows.

(1) A new design idea for image inpainting based on decreasing receptive field is proposed. To address the degradation problem in multi-stage inpainting networks, degression receptive field network (DRFNet) is proposed, which can significantly improve the quality of image inpainting. The proposed model takes receptive field as perspective and consists of five sub-networks with decreasing receptive fields. Quantitative and qualitative experiments show that DRFNet reaches state-of-the-art performance on three benchmark datasets: CelebA-HQ, Paris StreetView, and Places2.
(2) An easy-to-use TransConv module is designed. To overcome the problem of local-pixel influence in convolution, TransConv module is constructed that can improve the ability of the network to establish long-range dependencies. Unlike existing inpainting methods that use transformers, Transconv is designed to replace convolutions directly without requiring frequent feature and token conversions. Ablation experiments have demonstrated the effectiveness of Transconv.

2. Proposed method

2.1. Overall architecture

The receptive field refers to the range of pixels corresponding to neurons in a layer of network on input image, which is an essential attribute in deep neural networks. A large receptive field can reconstruct basic image structure from a small amount of high-dimensional information, and a smaller receptive field can better optimize image details by using the information of neighboring pixels. Therefore, DRFNet was designed based on the properties of receptive fields.

DRFNet generally follows coarse-to-fine design approach but still has significant differences compared with existing coarse-to-fine inpainting models, as shown in Fig. 5. Firstly, design ideas are different, existing multi-stage inpainting models are designed to realize coarse-to-fine by increasing network complexity, which may lead to degradation. DRFNet designs network based on the mechanism of receptive field in image inpainting and overcomes the degradation problem in existing coarse-to-fine network. Secondly, different from existing structures, DRFNet focuses on combined form of receptive fields and exploits the decreasing structure of receptive fields to realize inpainting from coarse-to-fine. Specifically, as shown in Fig. 6, DRFNet mainly consists of five blocks with decreasing receptive fields and selects a U-shaped network structure as first block; then, decreasing receptive field is achieved by gradually reducing the number of layers based on U-shaped block. Meanwhile, a TransConv module is designed and deployed in DRFNet to overcome the problem of local influence in convolution.

2.2. Degression receptive field network

This subsection introduces network structure of DRFNet in detail. As shown in Fig. 6, first original image and mask image are merged and fed into Block1. Then inpainted image of Block1 and mask are merged and input into Block2, and so on, until image output by Block5 is final result. The detailed structure of Block1 to Block5 is shown in Table 1, and the first five Conv modules in encoder of Block1 are replaced by TransConv. As shown in Table 1, continuously reduce pairs of Conv layers based on Block1 to reduce receptive field. For example, Block2 compared to Block1 receptive field drop of 382x382 by reducing one Conv layer and one DeConv layer. Table 2 shows that Conv module can be directly replaced with TransConv when parameters are consistent. DRFNet is described in detail as follows.

Specifically, given an input image $I_{input} \in \mathbb{R}^{C\times H\times W}$ and mask $M \in [0,1]^{H\times W}$ as initial input, formula is as follows:

$$I_{merged} = I_{input} \odot (1 - M) + M \tag{1}$$

$$I^{i}_{Output} = \text{Block}_i\left[\text{Cat}\left(I_{merged}, M\right)\right] \quad i = 1 \tag{2}$$

where I_{merged} is the input for Block_1, Block_i represents the *i-th* sub-

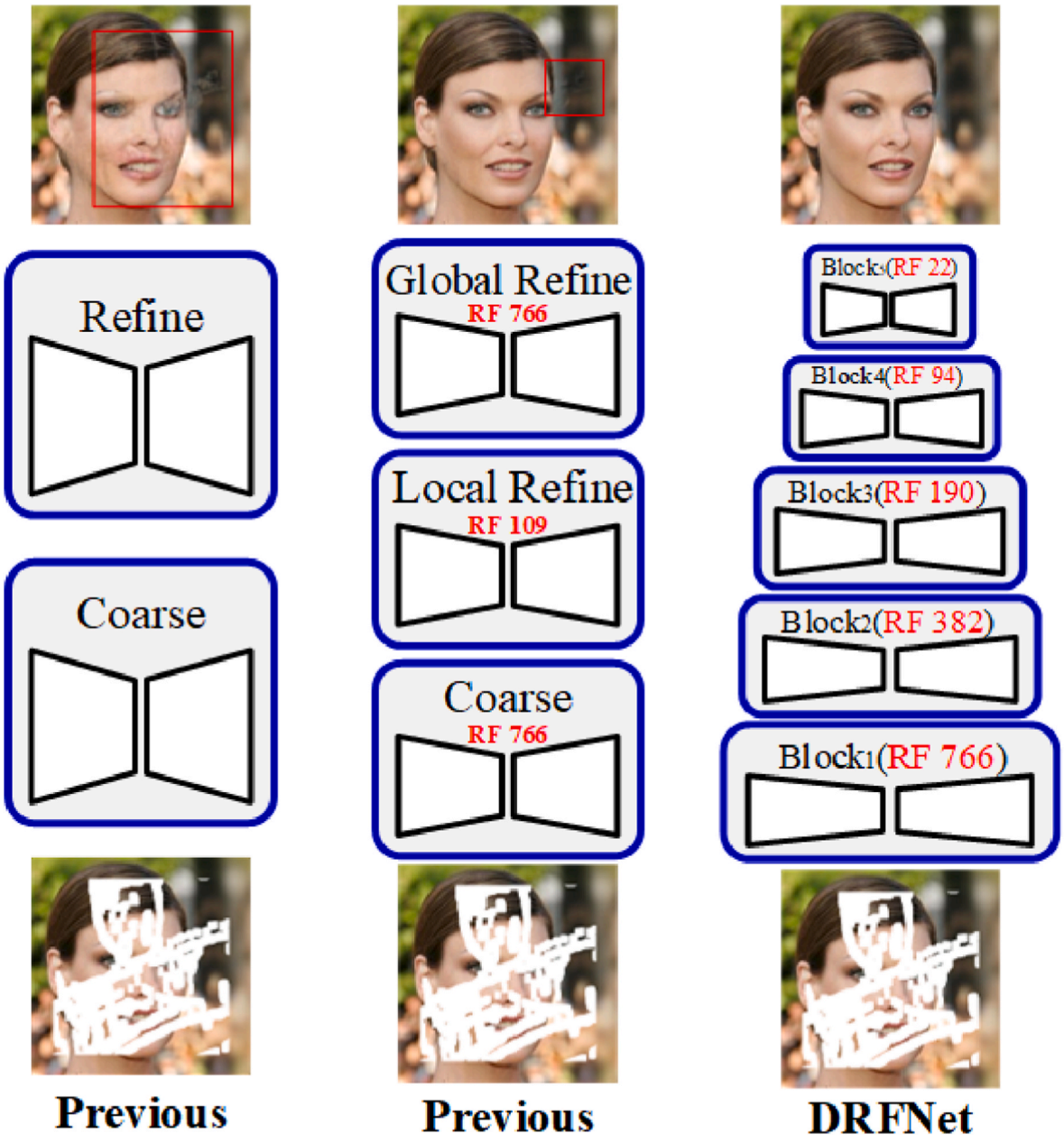


Fig. 5. The comparison diagram between DRFNet and existing coarse-to-fine methods. DRFNet aligns more with principles of receptive field application and differs significantly from existing method design paradigms. The total number of parameters for Block_3 to Block_5 was only 87% of global refine stage in LGNet (55.10 M).

network with different receptive fields in DRFNet, $I^{i}_{Output} \in \mathbb{R}^{C\times H\times W}$ represents output result of Block_i, $\odot$ represents point-wise multiplication, and operation $\text{Cat}(\cdot)$ denotes concatenate. The output I^{i}_{Output} of each block is calculated with original image I_{input} and mask M to obtain $I^{Block_i}_{merged} \in \mathbb{R}^{(C+1)\times H\times W}$.

$$I^{Block_i}_{merged} = I_{input} \odot (1-M) + I^{i}_{Output} \odot M \quad i \in \left[1, n\right) \tag{3}$$

Where $I^{Block_i}_{merged}$ denotes the result of merging repaired result I^{i}_{output} of each block with undamaged area of original image I_{iutput}.

Subsequently, $I^{Block_i}_{merged}$ and M are sent to next Block_{i+1} $i \in [1, n)$. The specific formula is as follows:

$$I^{i+1}_{Output} = \text{Block}_{i+1}\left[\text{Cat}\left(I^{Block_i}_{merged}, M\right)\right] \quad i \in \left[1, n\right) \tag{4}$$

Then I^{i+1}_{output} is sent back to (2) and (3). After passing through multiple sub-networks, final output I^{n}_{output} is obtained, which is the final inpainting image. For completeness, the inference procedures of DRFNet are summarized in **Algorithm 1**.

Algorithm 1: Inference procedure of DRFNet.
Initialization: Block_i is the i-th sub-network;
Input: input image $I_{input} \in \mathbb{R}^{C\times H\times W}$, mask $M \in [0,1]^{H\times W}$;
Output: final repaired image I^{n}_{output};
Inference: for *batchsize*← 1 ***to*** *total_sample* **do**
$I_{merged} \leftarrow I_{input} \odot (1-M) + M$//input merged image for Block_1
$I^{1}_{Output} \leftarrow \text{Block}_1\left[\text{Cat}(I_{merged}, M)\right]$//output of Block_1 **for** i←1 **to** n **do**//n is total number of Block
$I^{Block_i}_{merged} \leftarrow I_{input} \odot (1-M) + I^{i}_{Output} \odot M$//input merged image for Block_{i+1}
$I^{i+1}_{Output} \leftarrow \text{Block}_{i+1}\left[\text{Cat}\left(I^{Block_i}_{merged}, M\right)\right]$//output of Block_{i+1} **end for**
end for

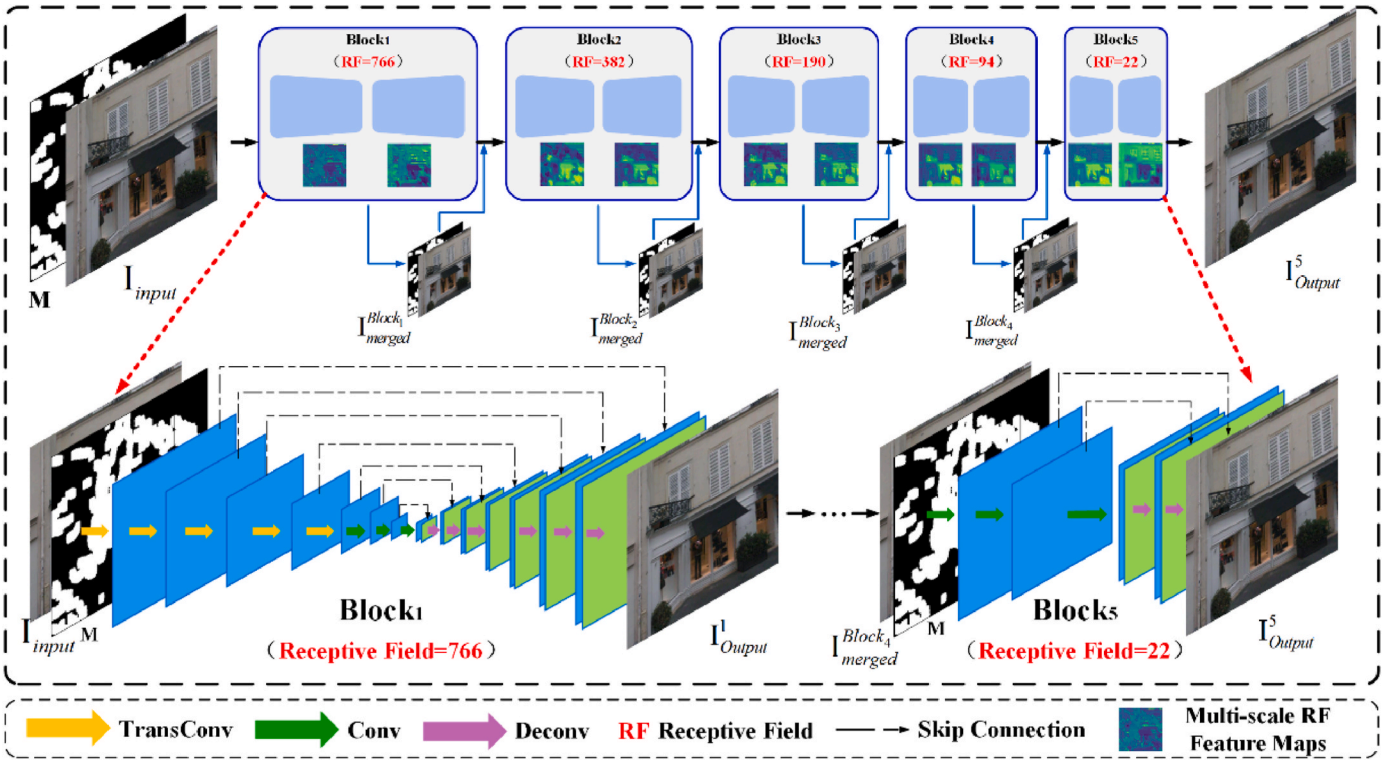


Fig. 6. Architecture of the proposed Degression Receptive Field Network (DRFNet). The upper half of figure shows network overview and lower half displays block details.

Table 1
The structure of DRFNet. RF denotes Receptive Field, Conv represents convolution operation, DeConv denotes deconvolution operation, and TransConv is the proposed module.

Stage	Layers (encoder)	Layers (decoder)	RF
$Block_1$	TransConv1, TransConv2, TransConv3, TransConv4, TransConv5, Conv6,Conv7,Conv8	DeConv_1, DeConv_2, DeConv_3, DeConv_4, DeConv_5, DeConv_6, DeConv_7, DeConv_8	766x766
$Block_2$	Conv1,Conv2,Conv3, Conv4,Conv5,Conv6, Conv7	DeConv_1, DeConv_2, DeConv_3, DeConv_4, DeConv_5, DeConv_6, DeConv_7	382x382
$Block_3$	Conv1,Conv2,Conv3, Conv4,Conv5,Conv6	DeConv_1, DeConv_2, DeConv_3, DeConv_4, DeConv_5, DeConv_6,	190x190
$Block_4$	Conv1,Conv2,Conv3, Conv4,Conv5	DeConv_1, DeConv_2, DeConv_3, DeConv_4, DeConv_5	94x94
$Block_5$	Conv1,Conv2,Conv3	DeConv_1, DeConv_2, DeConv_3	22x22

Table 2
The specific parameters of Conv, DeConv and TransConv layers. K denotes Kernel, S denotes Stride, and P denotes Padding.

Layer_name	Feature_size (in, out)	Channal_num (in, out)	K	S	P
Conv1	(256,128)	(4,64)	4	2	1
Conv2	(128,64)	(64,128)	4	2	1
Conv3	(64,32)	(128,256)	4	2	1
Conv4	(32,16)	(256,512)	4	2	1
Conv5	(16,8)	(512,512)	4	2	1
Conv6	(8,4)	(512,512)	4	2	1
Conv7	(4,2)	(512,512)	4	2	1
Conv8	(2,1)	(512,512)	4	2	1
DeConv_1	(1,2)	(512,512)	4	2	1
DeConv_2	(2,4)	(1024,512)	4	2	1
DeConv_3	(4,8)	(1024,512)	4	2	1
DeConv_4	(8,16)	(1024,512)	4	2	1
DeConv_5	(16,32)	(1024,256)	4	2	1
DeConv_6	(32,64)	(512,128)	4	2	1
DeConv_7	(64,128)	(256,64)	4	2	1
DeConv_8	(128,256)	(128,3)	4	2	1
TransConv1	(256,128)	(4,64)	4	2	1
TransConv2	(128,64)	(64,128)	4	2	1
TransConv3	(64,32)	(128,256)	4	2	1
TransConv4	(32,16)	(256,512)	4	2	1
TransConv5	(16,8)	(512,512)	4	2	1

2.3. TransConv

To address the problem of local-pixels influence in pure convolutional networks, a TransConv module is proposed that combines the advantages of convolution and transformer. TranConv has significant differences with existing transformer-based image inpainting models. Firstly, design idea is different. The current inpainting model is designed with a transformer as the encoder of generative network, using a convolutional neural network to convert input images into feature maps with smaller sizes and then feeding feature maps into transformer encoder for encoding, and finally decoding the encoded results using decoder, as shown in Previsou1 and Previsou2 of Fig. 7. In contrast, TransConv is designed for ease-of-use in mind, which allows direct replace of convolution. Secondly, specific structure is different. The current approach usually introduces a transformer encoder directly, which requires frequent conversions between feature maps and tokens in network. For example, in patch merging operation of swin transformer block in Previsou2 of Fig. 7. In order to reduce the size of feature maps, needs to convert tokens to feature maps, perform a convolution operation, and then convert feature maps into tokens. Meanwhile, the input of swin transformer block is required to be tokens, which means that feature map needs to be pre-processed when using swin transformer block. However, TransConv performs a single convolution operation on input feature map, which does not require frequent feature map and

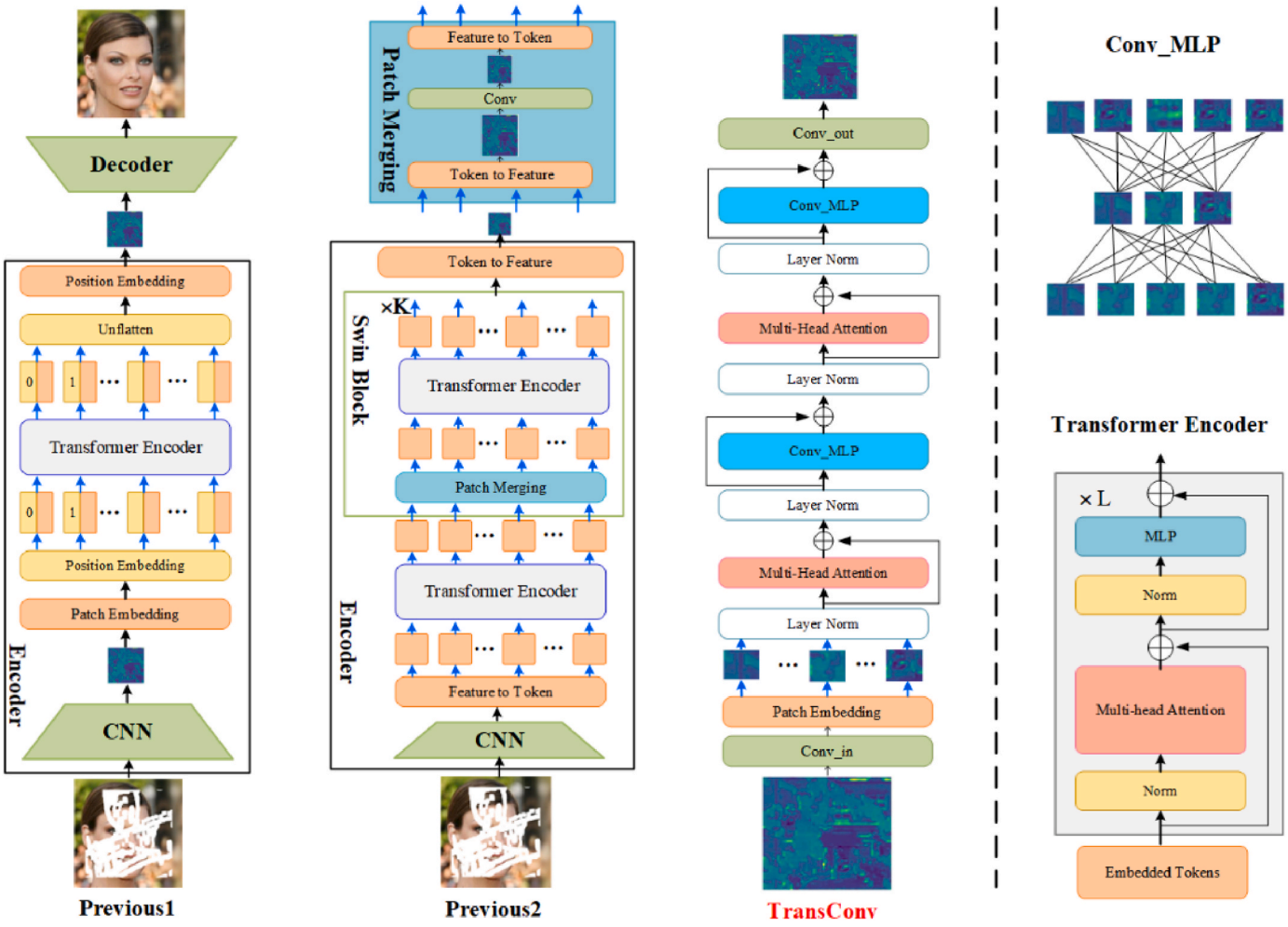


Fig. 7. The comparison diagram between TransConv and existing image inpainting methods that use transformer block.

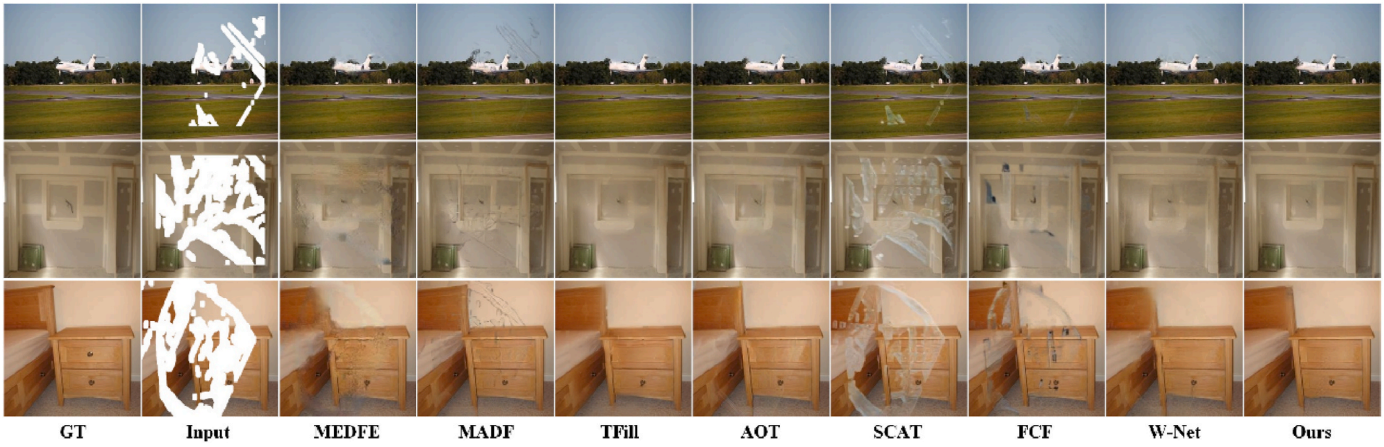


Fig. 8. Qualitative comparisons of the proposed method DRFNet with MEDFE, MADF, TFill, AOT, SCAT, FCF, and W-Net on Places2 Dataset. GT represents Ground Truth. Zoom in for more details. More results are available in the supplementary material.

token conversion. In addition, the transformer encoder is decomposed, reorganized, and optimized during the design of TransConv, which eliminates the depth setting of transformer block and reduces linear transformation between features. TransConv is described in detail as follows.

Let $F_{input} \in \mathbb{R}^{b\times c1\times h\times w}$ be a small batch of input feature maps in $Block_1$, where b, c_1, h, w represent batch size, number of channels, width, and height, respectively. In order to preserve inductive bias of convolution and reduce calculation quantity of multi-head attention (MHA), a convolution operation is first performed on input feature map to reduce the size of feature map and obtain output feature $F_{C\_in} \in \mathbb{R}^{b\times c'\times h'\times w'}$.

$$F_{C\_in} = \mathrm{Conv\_in}(F_{input}) \tag{5}$$

Fig. 9. Qualitative comparisons of the proposed method DRFNet with MADF, TFill, MAT, LGNet, AOT, FCF, and W-Net on CelebA-HQ Dataset. GT represents Ground Truth. Zoom in for more details. More results are available in the supplementary material.

Fig. 10. Qualitative comparisons of the proposed method DRFNet with MEDFE, MADF, W-Net, and Ours on Paris StreetView Dataset. GT represents Ground Truth. Zoom in for more details. More results are available in the supplementary material.

where c', h', and w' represent the number of output channels, length, and width after Conv_in$(\cdot)$ operation, respectively.

Secondly, for subsequent calculation of MHA, Patch_Embedding$(\cdot)$ operation is used to cut $F_{C\_in} \in \mathbb{R}^{b\times c'\times h'\times w'}$ into pieces and obtain $F_{PE\_out} \in \mathbb{R}^{b\times n\times (d\times p^2)}$.

$$F_{PE\_out} = \text{Patch\_Embedding}(F_{C\_in}) \tag{6}$$

where p represents the size of each image patch, $n = h'w'/p^2$ is the number of embedded patches, and d is hidden layer size.

Thirdly, in order to obtain attention feature $F_{M\_out} \in \mathbb{R}^{b\times n\times (d\times p^2)}$, $F_{PE\_out} \in \mathbb{R}^{b\times n\times (d\times p^2)}$ is sent to MHA for self-attention calculation after layer normalization.

$$F_{M\_out} = \text{MHA}[\text{LN}(F_{PE\_out})] + F_{PE\_out} \tag{7}$$

where LN$(\cdot)$ operation represents layer normalization.

Fourth, to adjust image characteristics, Conv_MLP$(\cdot)$ is designed to remap the features by exchanging the linear transformation with feature-level operations to obtain $F_{CM\_out} \in \mathbb{R}^{b\times n\times (d\times p^2)}$, as shown in Fig. 7.

$$F_{CM\_out} = \text{Conv\_MLP}[\text{LN}(F_{M\_out})] + F_{M\_out} \tag{8}$$

Finally, to ensure ease-of-use of TransConv, features are restored to $F_{C\_out} \in \mathbb{R}^{b\times c'\times h'\times w'}$ by Conv_out$(\cdot)$.

$$F_{C\_out} = \text{Conv\_out}(F_{CM\_out}) \tag{9}$$

2.4. Loss functions

For better training of model, DRFNet adopts composite loss function during training process. The settings of loss function are described below.

Firstly, in order to optimize training process and reduce training time, Block$_i$ $i \in [1, n]$ adopts the weighted L1 loss $\ell_{vh}^{block_i}$ for pixel-level reconstruction, formulas are as follows:

Fig. 11. The subjective verification experiment of TransConv. From left to right are Ground Truth, Masked Image, DRFNet_C, DRFNet_T, and DRFNet. Zoom in for more details.

$$\ell_{valid}^{block_i} = \frac{1}{\mathrm{sum}(1-\mathrm{M})}\left\| \left(\mathrm{I}_{Output}^{i} - \mathrm{I}_{input}\right) \odot (1-\mathrm{M}) \right\|_1 \quad i \in [1,n] \tag{10}$$

$$\ell_{hole}^{block_i} = \frac{1}{\mathrm{sum}(\mathrm{M})}\left\| \left(\mathrm{I}_{Output}^{i} - \mathrm{I}_{input}\right) \odot \mathrm{M} \right\|_1 \quad i \in [1,n] \tag{11}$$

$$\ell_{vh}^{block_i} = \ell_{valid}^{block_i} + \lambda_h \cdot \ell_{hole}^{block_i} \quad i \in [1,n] \tag{12}$$

Secondly, to enhance the recovery of structural and textural details, Block_n incorporates perceptual loss (Johnson et al., 2016) and style loss (Gatys et al., 2016), as recommended by literature (Liu et al., 2018; Quan et al., 2022; Zhu et al., 2021). These losses are instrumental in capturing high-level semantics and overall visual appeal of repaired images. Additionally, the total variation (TV) loss (Liu et al., 2018) is applied as smoothing penalty before applying perceptual and style losses. The loss function of generator is defined as:

$$\begin{aligned} \ell_G = \sum_{i=1}^{n} \ell_{vh}^{block_i} + \lambda_{tv} \cdot \ell_{tv}^{block_n} + \lambda_{perceptual} \cdot \ell_{perceptual}^{block_n} \\ + \lambda_{style} \cdot \ell_{style}^{block_n} \quad i \in [1,n] \end{aligned} \tag{13}$$

Adversarial training is almost essential to improve quality of generated images. DRFNet also introduces adversarial loss during training, same as works (Chen et al., 2023b; Kumar et al., 2024; Li et al., 2023; Quan et al., 2022). The corresponding loss function of discriminator is as follows:

$$\begin{aligned} \ell_D = \min_G \max_D \mathbb{E}_{\mathrm{I}_{input} \sim \mathbb{P}_r}\left[logD\left(\mathrm{I}_{input}\right)\right] - \\ \mathbb{E}_{G \sim \mathbb{P}_g}\left[log\left(D\left(G\left(\mathrm{I}_{merged}^{Block_4}, \mathrm{M}\right)\right)\right)\right] \end{aligned} \tag{14}$$

To sum up, the proposed inpainting model DRFNet is trained in end-to-end manner, and final total loss function ℓ can be formulated as:

$$\ell = \ell_G + \ell_D \tag{15}$$

3. Experiments

In this section, experimental setup is first introduced, which mainly includes dataset, introduction of comparison methods, evaluation indicators, and experimental details. Subsequently, the effectiveness of proposed model DRFNet was verified through comparative experiments with state-of-the-art methods, and further analysis of DRFNet was conducted through ablation experiments. Finally, application cases of DRFNet in real scenarios are demonstrated.

3.1. Experimental settings

1) ***Datasets*:** DRFNet was evaluated by using three benchmark datasets in image inpainting.
   - Paris StreetView (Doersch et al., 2015): A dataset of street view images. It contains urban buildings, trees, streets, skies, etc. collected from Paris, France. The original division of 14,900 images was taken as training set and 100 images as test set.

- CelebA-HQ (Liu et al., 2015): A high-resolution face dataset. The CelebA-HQ is a dataset containing 30,000 high-resolution face images extracted from CelebA dataset. 2,000 images were randomly selected for testing, and the remaining 28,000 images were for training.
- Places2 (Zhou et al., 2017): A large-scale scene recognition dataset. This part follows LGNet (Quan et al., 2022) for dataset selection. Specifically, 2000 images were randomly selected from first 20 categories for training and 100 images for testing. Overall, 40,000 images are taken as training set and 2000 images as test set. Please note that training and testing sets are independent of each other.

Quick Draw Irregular Mask Dataset (QD-IMD) (Iskakov, 2021) was adopted as mask during training. An irregular mask (Liu et al., 2018), commonly applied in image inpainting, was implemented for testing. Irregular mask contains six groups of different mask ratios (0.01, 0.1], (0.1, 0.2], (0.2, 0.3], (0.3, 0.4], (0.4, 0.5], (0.5, 0.6], and each group contains 2000 images. In the experiment, the sizes of all images and masks have been resized to 256x256.

2) ***Comparison methods*:** Nine state-of-the-art (SOTA) models were selected for comparative experiments, including seven models derived from prestigious computer vision conferences and journals for 2022 and 2023, respectively.
   - MEDFE (Liu et al., 2020) refers to "An image inpainting network with mutual codecs".
   - MADF (Zhu et al., 2021) refers to "A U-shaped image inpainting network with mask-aware dynamic filtering module and point normalization".
   - TFill (Zheng et al., 2022) refers to "A coarse-to-fine two-stage image inpainting network that introduces transformer to image inpainting for the first time".
   - MAT (Li et al., 2022) refers to "A transformer-based coarse-to-fine image inpainting network".
   - LGNet (Quan et al., 2022) refers to "A three-stage image inpainting framework with local and global refinement subnetworks".
   - SCAT(Zuo et al., 2023) refers to "A U-shaped image inpainting network with segmentation confusion adversarial training".
   - AOT (Zeng et al., 2023) refers to "An image inpainting network with aggregated contextual-transformation block".
   - FCF(Jain et al., 2023) refers to "An image inpainting network with Fourier Coarse-to-Fine generator".
   - W-Net (Zhang et al., 2023a) refers to "A two-stage network with coarse and refined structures derived at each stage".
3) ***Evaluation metrics*:** L1 error, Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) (Wang et al., 2004), Fréchet Inception Distance (FID) (Heusel et al., 2017), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) were applied to evaluate experimental results, which are commonly used in image inpainting tasks.

L1 error: L1 error measures the average absolute difference between predicted value and true value, also known as the Manhattan Distance or Sum of Absolute Errors.

SSIM: a measure of the similarity of two images that takes into account the structure, luminance, and contrast information of the image.

PSNR: a measure of image reconstruction quality, which is evaluated by comparing the mean square error between the original image and the reconstructed image.

FID: a metric for evaluating the performance of generative model that assesses the quality of a generated image by comparing the distribution of generated image with real image in feature space.

LPIPS: an image similarity measure based on deep learning. LPIPS employs a pre-trained convolutional neural network to extract features, which are then used to calculate the perceptual similarity between features.

L1 error, PSNR, and SSIM are pixel-level evaluation metrics. FID and LPIPS are perception-based evaluation metrics whose evaluation reisisults are more in line with human visual perception. Note that larger is better for PSNR, and SSIM, while smaller is better for FID and LPIPS.

4) ***Implementation details*:** All experiments were performed on Pytorch v1.10.0 and CUDA v11.3 with a single NVIDIA GeForce RTX 3090 GPU. DRFNet was optimized by Adam Optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning strategy for DRFNet is set with epochs = 200 through convergence training analysis. Specifically, initial learning rate is set to 0.0002 for the first 100 epochs, and learning rate decays linearly to 0 for the subsequent 100 epochs. The number of blocks n = 5. And parameter $p = 16$ and $d = 512$. Parameters of loss function $\lambda_h = 6$, $\lambda_{tv} = 0.1$, $\lambda_{perceptual} = 0.05$, and $\lambda_{style} = 120$ validated in references (Quan et al., 2022) are directly adopted.

3.2. Comparisons with state-of-the-art methods

DRFNet was compared both quantitatively and qualitatively with nine SOTA methods to demonstrate the effectiveness and advanced. At the same time, all models were tested using official code and pre-trained model to ensure fairness of comparison experiments. The test set and masks are same for each model.

Please note that: In order to further validate the authenticity of the data, the evaluation code (Zheng et al., 2022), randomly selected test sets and masks used in this work already available on GitHub.

1) ***Quantitative comparisons*:** The results of quantitative comparison between DRFNet and SOTA models on Places2, CelebA-HQ, and Paris StreetView datasets are shown in Tables 3–5 The comprehensive inpainting metrics of DRFNet are significantly outperforming than other SOTA methods, as average values of four metrics have achieved best results. Only a few metrics of DRFNet did not achieve best results, but almost all of them are still in second place and only marginally different from the best, which is still competitive (e.g., PSNR metrics of DRFNet on Places2 dataset only rank second under 50%–60% mask ratio, which is 0.4% different from best method W-Net. However, the average PSNR of DRFNet is 4.6% better than W-Net).
2) ***Qualitative comparisons*:** DRFNet is competitive based on results of qualitative comparison experiments on three datasets Places2, ClelebA-HQ, and Paris StreetView, as shown in Figs. 8–10 DRFNet has better structure reconstruction ability and superior artifact removal ability. For example, by comparing the background in the third column of Fig. 9, it can be seen more intuitively that DRFNet has a better artifact removal ability. From structural repair result of airplane in first row of images in Fig. 8, DRFNet has superior inpainting ability. The inpainting ability of DRFNet is due to stepwise refinement repair ability of decreasing receptive field network to damaged area and long-range dependency establishment ability of TransConv.

3.3. Ablation studies

1) ***TransConv*:** This part was performed on DRFNet to evaluate performance of TransConv, and results are shown in Table 6. Based on the data in Table 6, the values of various metrics for DRFNet with TransConv comprehensively outperform DRFNet with only using vanilla convolution or transformer. Meanwhile, to intuitively prove advantages of TransConv, four sets of hand-drawn masks were subjectively experimented on CelebA-HQ, and results are shown in Fig. 11. It can be visualized that TransConv has better global contextual connectivity and long-range dependency establishment capabilities by comparing columns 3 to 5 in Fig. 11. For example, in the first row and third column, DRFNet_C failed to inpaint structural

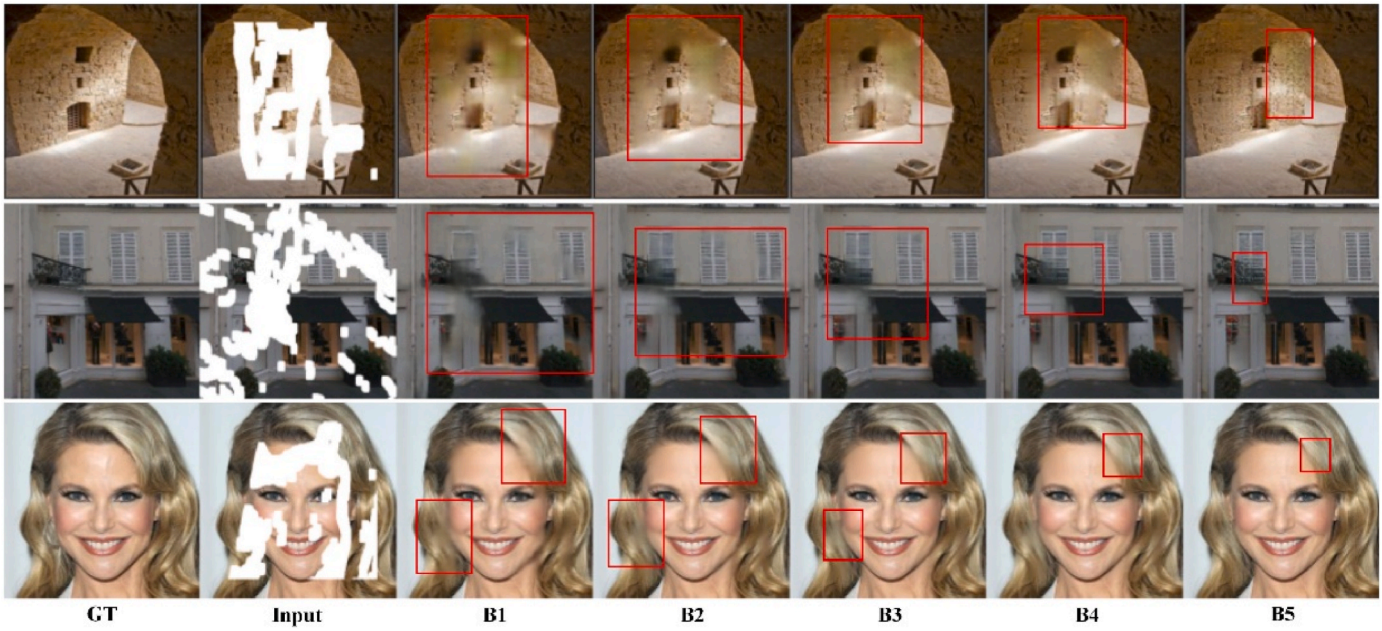


Fig. 12. The decomposition experiments of DRFNet. Zoom in for more details.

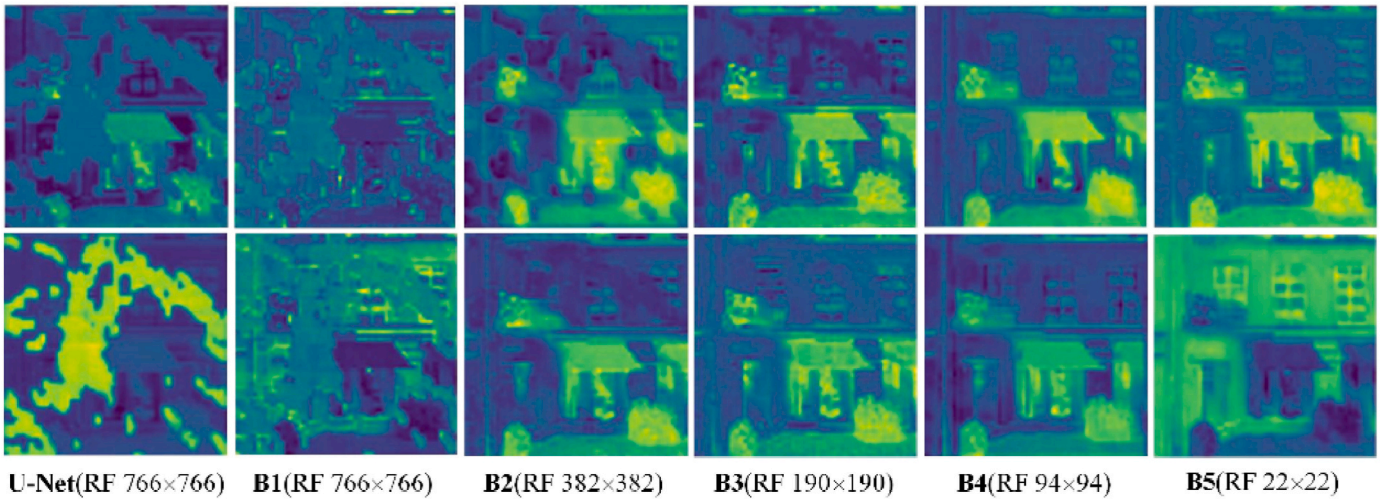


Fig. 13. Intermediate feature maps of UNet and DRFNet, where the receptive field for each network is indicated in parentheses. B1-5 represent the Block1-5 in DRFNet. Zoom in for more details.

and detailed information; in the fourth column, DRFNet_T has inpainted general contour, but due to the nature of transformer, a block effect is present in blue area at the top right; in last column, DRFNet's inpaint results are more natural in contour, and the consistency of blue area at the top right has improved.

2) ***Network design*:** The network design of DRFNet was verified and evaluated by visualizing output results of each block step by step, and objective data are shown in Table 7, subjective visual comparison is shown in Fig. 12. Please note that receptive fields of B_1 to B_5 are 766, 382, 190, 94, and 22, respectively. Observing red box in Fig. 12 with decreases of receptive field more and more details are gradually perfected, which is completely consistent with trend of FID in Table 7. In summary, Table 7 and Fig. 12 fully support network design: the decreasing structure of receptive field from large to small can gradually refine inpainting result.

Additionally, to intuitively verify the effectiveness and rationality of DRFNet design, this part extracted and visualized the intermediate feature maps of DRFNet and a single U-Net for analysis. Specifically, to present the details within feature maps, intermediate feature maps of size 64x64 were extracted from a single U-Net and each sub-network within DRFNet for analysis. The results are depicted in Fig. 13. Firstly, by comparing the feature maps of B1-B5, it can be observed that as the receptive field of sub-networks in DRFNet decreases, the details in feature maps extracted at each block are progressively filled in. This is consistent with visualized results in Fig. 12, thereby corroborating effectiveness of DRFNet with decreasing receptive field. Secondly, by comparing the feature maps extracted by U-Net with corresponding feature maps from each subnetwork of DRFNet, it is evident that feature maps from a single U-Net network exhibit clear masking artifacts. In contrast, feature maps extracted from cascaded networks with decreasing receptive fields show a gradual weakening of these artifacts. This further validates the rationality of DRFNet network architecture.

3) ***Loss function analysis:*** DRFNet mainly focuses on the design of decreasing receptive field network, the loss function directly adopts

Table 3
Quantitative comparisons on **Places2** with irregular masks. The two best scores are indicated by Red and Blue fonts. Please note that ⬆ Higher is better And ⬇ Lower is better.

Metric	Method	Mask Ratio						Average
		1-10%	10-20%	20-30%	30-40%	40-50%	50-60%	
PSNR⬆	MEDFE	33.67	27.77	24.41	22.07	20.28	17.90	24.35
	MADF	33.64	28.23	25.20	23.10	21.41	19.02	25.10
	TFill	33.48	27.86	24.85	22.83	21.21	19.06	24.88
	AOT	31.78	28.78	25.64	23.26	21.37	18.75	24.93
	SCAT	31.27	26.21	23.46	21.57	20.05	18.08	23.44
	FCF	34.17	28.06	24.69	22.34	20.54	17.93	24.62
	W-Net	33.73	28.57	25.50	23.39	21.70	19.39	25.38
	Ours	36.78	30.32	26.66	24.08	22.13	19.31	26.55
SSIM⬆	MEDFE	0.973	0.925	0.861	0.789	0.711	0.602	0.810
	MADF	0.966	0.915	0.855	0.791	0.721	0.624	0.812
	TFill	0.974	0.929	0.873	0.814	0.748	0.653	0.832
	AOT	0.964	0.934	0.882	0.820	0.749	0.640	0.831
	SCAT	0.959	0.902	0.837	0.768	0.695	0.602	0.794
	FCF	0.974	0.929	0.871	0.806	0.735	0.627	0.824
	W-Net	0.972	0.929	0.875	0.815	0.747	0.651	0.832
	Ours	0.984	0.952	0.904	0.847	0.781	0.665	0.855
LPIPS⬇	MEDFE	0.023	0.063	0.116	0.175	0.238	0.328	0.157
	MADF	0.031	0.069	0.111	0.152	0.198	0.270	0.139
	TFill	0.019	0.050	0.090	0.133	0.182	0.263	0.123
	AOT	0.039	0.048	0.081	0.121	0.167	0.246	0.117
	SCAT	0.037	0.081	0.127	0.172	0.220	0.286	0.154
	FCF	0.018	0.050	0.088	0.130	0.177	0.254	0.120
	W-Net	0.022	0.053	0.093	0.138	0.190	0.278	0.129
	Ours	0.011	0.034	0.067	0.109	0.158	0.256	0.106
FID⬇	MEDFE	4.74	12.37	23.12	35.28	50.70	72.00	33.03
	MADF	4.81	10.18	16.60	22.20	28.89	39.91	20.43
	TFill	3.63	8.86	14.88	20.80	27.86	40.24	19.38
	AOT	7.05	8.38	13.49	19.38	26.40	40.68	19.23
	SCAT	6.35	13.51	21.45	28.90	37.41	48.95	26.09
	FCF	3.65	8.44	14.19	19.33	25.03	33.62	17.38
	W-Net	4.64	10.49	18.48	27.28	38.94	59.06	26.48
	Ours	2.20	6.04	11.41	17.68	25.92	47.32	18.43
ℓ_1(%) ⬇	MEDFE	0.57	1.30	2.26	3.38	4.64	6.92	3.18
	MADF	0.42	1.10	1.95	2.89	3.94	5.83	2.69
	TFill	0.50	1.15	1.99	2.90	3.92	5.70	2.69
	AOT	0.60	1.07	1.84	2.80	3.92	6.01	2.71
	SCAT	0.65	1.54	2.60	3.73	4.96	6.90	3.40
	FCF	0.39	1.08	1.99	3.05	4.25	6.52	2.88
	W-Net	0.64	1.26	2.05	2.94	3.97	5.78	2.77
	Ours	0.37	0.87	1.58	2.45	3.46	5.53	2.38

composite loss function commonly used in image inpainting according to reference (Phutke and Murala, 2023; Quan et al., 2022; Zhu et al., 2021). However, it is still essential to analyzing the actual effect of loss function in DRFNet. Therefore, ablation experiments were conducted on L1 loss, TV loss, Perceptual loss, and Style loss using Paris StreetView dataset. The experimental results are shown in Table 8 and Fig. 14.

An analysis of the data presented in Table 8 within DRFNet network indicates that L1 loss, TV loss, and Perceptual loss show positive effects on improving PSNR and SSIM. Nevertheless, their efficacy in augmenting visual-related metrics appears to be constrained. In contrast, Style loss significantly reduces LPIPS and FID metrics but sacrifices the accuracy of pixel-level metrics. Further analysis of Fig. 14 reveals the following: Firstly, L1 loss serving as a pixel-level reconstruction metric

Table 4
Quantitative comparisons on **CelebA-HQ** with irregular masks. The two best scores are indicated by Red and Blue fonts. Please note that ↑ Higher is better And ↓ Lower is better.

Metric	Method	Mask Ratio						Average
		1-10%	10-20%	20-30%	30-40%	40-50%	50-60%	
PSNR↑	MADF	37.25	31.78	28.51	25.96	23.97	20.72	28.03
	TFill	36.47	31.00	27.94	25.66	23.97	21.47	27.75
	MAT	38.62	32.65	29.30	26.72	24.80	21.81	28.98
	LGNet	38.08	33.04	29.97	27.60	25.80	22.97	29.58
	AOT	34.28	29.10	25.62	23.16	21.34	18.62	25.35
	FCF	37.48	32.03	28.84	26.43	24.67	22.04	28.58
	W-Net	36.68	32.04	29.20	26.95	25.17	22.59	28.77
	Ours	40.79	34.70	30.93	28.05	25.91	22.47	30.48
SSIM↑	MADF	0.976	0.941	0.899	0.850	0.795	0.697	0.859
	TFill	0.979	0.944	0.904	0.858	0.807	0.726	0.870
	MAT	0.985	0.958	0.921	0.877	0.827	0.739	0.885
	LGNet	0.981	0.955	0.921	0.882	0.838	0.763	0.890
	AOT	0.969	0.926	0.868	0.805	0.738	0.643	0.825
	FCF	0.980	0.948	0.908	0.863	0.814	0.739	0.875
	W-Net	0.978	0.949	0.914	0.873	0.828	0.758	0.883
	Ours	0.990	0.971	0.942	0.905	0.861	0.775	0.907
LPIPS↓	MADF	0.019	0.043	0.070	0.100	0.136	0.205	0.095
	TFill	0.011	0.028	0.048	0.071	0.097	0.138	0.065
	MAT	0.007	0.019	0.036	0.056	0.079	0.123	0.053
	LGNet	0.012	0.026	0.043	0.062	0.083	0.121	0.058
	AOT	0.023	0.048	0.084	0.122	0.164	0.234	0.113
	FCF	0.011	0.027	0.046	0.067	0.089	0.125	0.061
	W-Net	0.013	0.029	0.048	0.071	0.096	0.138	0.066
	Ours	0.005	0.015	0.030	0.049	0.073	0.124	0.049
FID↓	MADF	2.89	5.99	10.29	16.66	24.45	48.46	18.12
	TFill	1.61	3.58	5.98	8.41	11.49	16.09	7.86
	MAT	1.12	2.62	4.46	6.81	9.12	12.71	6.14
	LGNet	2.07	3.65	5.71	7.95	10.53	15.25	7.53
	AOT	4.17	7.00	12.68	20.29	30.90	57.53	22.10
	FCF	2.00	3.92	5.83	7.71	9.37	10.90	6.62
	W-Net	2.37	4.04	5.93	8.09	10.32	13.09	7.31
	Ours	0.80	2.06	3.88	6.15	9.04	15.12	6.17
ℓ_1(%) ↓	MADF	0.26	0.68	1.23	1.90	2.69	4.48	1.87
	TFill	0.39	0.82	1.38	2.04	2.77	4.13	1.92
	MAT	0.22	0.60	1.11	1.74	2.45	3.93	1.68
	LGNet	0.34	0.68	1.13	1.66	2.27	3.48	1.59
	AOT	0.50	1.11	1.98	2.99	4.12	6.29	2.83
	FCF	0.26	0.69	1.25	1.89	2.60	3.88	1.76
	W-Net	0.52	0.90	1.37	1.94	2.59	3.79	1.85
	Ours	0.27	0.55	0.95	1.49	2.11	3.56	1.49

can reconstruct the overall outline of masked area, but with some loss of detail. Secondly, TV loss is added as a smoothing penalty term to achieve smoother inpainting results, as evidenced by comparing results in the third and fourth columns of Fig. 14. Thirdly, comparing the fourth and fifth columns in Fig. 14, it can be observed that Perceptual loss can further enhance semantic consistency of repaired area based on L1 loss and TV loss, but improvement is limited. Finally, the introduction of Style loss leads to a significant improvement in texture detail repair of masked areas.

In summary, although the composite loss function employed by DRFNet has achieved satisfactory comprehensive repair results, there remains scope for further optimization of the loss function in terms of subjective results and objective metrics, indicating potential for further research.

Table 5
Quantitative comparisons on **Paris StreetView** with irregular masks. The two best scores are indicated by Red and Blue fonts. Please note that ↑ Higher is better And ↓ Lower is better.

Metric	Method	Mask Ratio						Average
		1-10%	10-20%	20-30%	30-40%	40-50%	50-60%	
PSNR↑	MEDFE	36.86	31.25	27.92	25.72	23.38	20.92	27.67
	MADF	35.45	30.46	27.59	25.90	24.00	21.90	27.55
	W-Net	35.41	30.94	28.13	26.38	24.48	22.44	27.96
	Ours	38.04	32.41	29.03	26.82	24.58	22.04	28.82
SSIM↑	MEDFE	0.982	0.946	0.896	0.844	0.766	0.659	0.849
	MADF	0.972	0.926	0.876	0.826	0.764	0.677	0.840
	W-Net	0.975	0.939	0.893	0.848	0.786	0.702	0.857
	Ours	0.985	0.956	0.913	0.868	0.803	0.702	0.871
LPIPS↓	MEDFE	0.013	0.040	0.076	0.112	0.165	0.241	0.108
	MADF	0.022	0.057	0.090	0.123	0.163	0.221	0.112
	W-Net	0.020	0.046	0.077	0.110	0.155	0.225	0.106
	Ours	0.010	0.031	0.061	0.094	0.140	0.221	0.093
FID↓	MEDFE	8.29	18.94	31.62	44.94	61.67	80.61	41.01
	MADF	12.60	25.61	36.42	46.59	60.08	69.76	41.84
	W-Net	14.05	22.99	33.30	45.21	61.77	76.83	42.36
	Ours	6.04	13.76	23.51	35.61	51.55	71.15	33.60
ℓ_1(%) ↓	MEDFE	0.39	0.87	1.53	2.20	3.26	4.95	2.20
	MADF	0.34	0.91	1.57	2.17	3.03	4.36	2.06
	W-Net	0.33	0.82	1.43	2.01	2.84	4.11	1.92
	Ours	0.35	0.76	1.33	1.90	2.77	4.27	1.90

Table 6
The verification experiment of TransConv. DRFNet_C represents DRFNet only using convolution, DRFNet_T represents the DRFNet network only using transformer, and DRFNet represents the DRFNet network after adding TransConv, which is main network. Please note that ↑ Higher is better And ↓ Lower is better.

Metric	Stage	Paris StreetView	CelebA-HQ	Places2
PSNR↑	DRFNet_C	28.18	29.77	25.75
	DRFNet_T	28.20	30.32	26.14
	DRFNet	28.82	30.48	26.55
SSIM↑	DRFNet_C	0.857	0.896	0.838
	DRFNet_T	0.856	0.905	0.843
	DRFNet	0.871	0.907	0.855
LPIPS↓	DRFNet_C	0.111	0.058	0.130
	DRFNet_T	0.109	0.052	0.124
	DRFNet	0.093	0.049	0.106
FID↓	DRFNet_C	41.08	8.38	26.83
	DRFNet_T	42.71	7.08	25.18
	DRFNet	33.60	6.17	18.43
ℓ_1(%)↓	DRFNet_C	2.10	1.70	2.66
	DRFNet_T	2.11	1.53	2.59
	DRFNet	1.90	1.49	2.38

Table 7
The **FID** comparisons of DRFNet decomposition experiments on three datasets. $B_{1\text{-}5}$ represent the intermediate results of $Block_{1\text{-}5}$ in DRFNet, respectively. Lower is Better.

	Mask Ratio	B_1	B_2	B_3	B_4	B_5
CelebA-HQ	1–10%	1.63	1.07	1.01	0.92	0.80
	10–20%	4.09	2.97	2.75	2.47	2.06
	20–30%	7.52	5.96	5.47	4.83	3.88
	30–40%	11.94	9.90	8.63	7.69	6.15
	40–50%	16.88	14.78	12.34	11.13	9.04
	50–60%	25.28	22.79	18.16	17.11	15.12
Places2	1–10%	4.43	3.04	2.65	2.46	2.20
	10–20%	12.66	8.95	7.59	6.86	6.04
	20–30%	25.06	17.99	14.72	12.90	11.41
	30–40%	41.63	31.05	24.02	20.19	17.68
	40–50%	62.56	48.53	36.53	28.99	25.92
	50–60%	95.68	84.79	67.44	50.71	47.32
Paris StreetView	1–10%	10.96	8.37	6.97	6.42	6.04
	10–20%	27.32	19.15	18.01	16.10	13.76
	20–30%	51.14	37.43	32.74	28.47	23.51
	30–40%	76.38	55.36	48.98	41.97	35.61
	40–50%	112.59	84.15	71.95	60.15	51.55
	50–60%	161.07	133.54	106.56	83.05	71.15

3.4. Computational complexity analysis

This section provides an in-depth analysis of computational complexity within DRFNet to enhance understanding of the network's operational intricacies. And the number of parameters, computational cost (Flops), training time (Train. time), and inference time (Infer. time) were selected as statistics. The experiment utilized Paris StreetView dataset as test data, conducting tests on a single NVIDIA GeForce RTX 3090 GPU and calculating the mean of results. Specific data are presented in Table 9.

Analysis of Tables 7 and 9 reveals that featuring a cascading reduction in receptive field did not lead to substantial increase in parameters during the design of DRFNet. Instead, Block3-5 has achieved a notable enhancement in network performance with a relatively small parameter increase. For example, in Table 7, with Paris StreetView dataset at mask ratio of 30–40%, Block3-5 achieved 35.7% improvement in the FID, while the total number of parameters was only 87% of the third-stage of LGNet (55.10 M).

Additionally, analyzing the data in Table 9 can also provide clear guidance for future model improvements. For example, the introduction of Transconv module has resulted in Block1 having more parameters than Blocks 2–5. Subsequent lightweight design optimizations of

Table 8
Results of quantitative analysis of the loss function. Per represents perceptual loss. Please note that ↑ Higher is better And ↓ Lower is better.

	Loss	1–10%	10–20%	20–30%	30–40%	40–50%	50–60%
PSNR↑	L1	38.39	32.88	29.54	27.42	25.17	22.57
	L1+ TV	38.40	32.87	29.59	27.45	25.16	22.65
	L1+ TV + Per	38.41	32.93	29.60	27.46	25.24	22.67
	L1+ TV + Per + Style	38.04	32.41	29.03	26.82	24.58	22.04
SSIM↑	L1	0.986	0.959	0.918	0.877	0.813	0.717
	L1+ TV	0.986	0.959	0.918	0.877	0.814	0.719
	L1+ TV + Per	0.986	0.959	0.919	0.877	0.816	0.720
	L1+ TV + Per + Style	0.985	0.956	0.913	0.868	0.803	0.702
LPIPS↓	L1	0.012	0.039	0.078	0.121	0.184	0.288
	L1+ TV	0.012	0.038	0.077	0.120	0.184	0.288
	L1+ TV + Per	0.012	0.039	0.078	0.122	0.185	0.292
	L1+ TV + Per + Style	0.009	0.031	0.061	0.094	0.140	0.221
FID↓	L1	6.31	16.44	30.15	45.97	69.28	110.26
	L1+ TV	6.24	16.03	28.77	44.02	67.56	106.73
	L1+ TV + Per	6.41	16.03	29.51	44.65	68.61	109.82
	L1+ TV + Per + Style	6.04	13.76	23.51	35.61	51.55	71.15
ℓ_1(%)↓	L1	0.34	0.72	1.24	1.76	2.58	4.03
	L1+ TV	0.34	0.72	1.24	1.76	2.58	4.01
	L1+ TV + Per	0.33	0.72	1.24	1.76	2.57	4.00
	L1+ TV + Per + Style	0.35	0.76	1.33	1.90	2.77	4.27

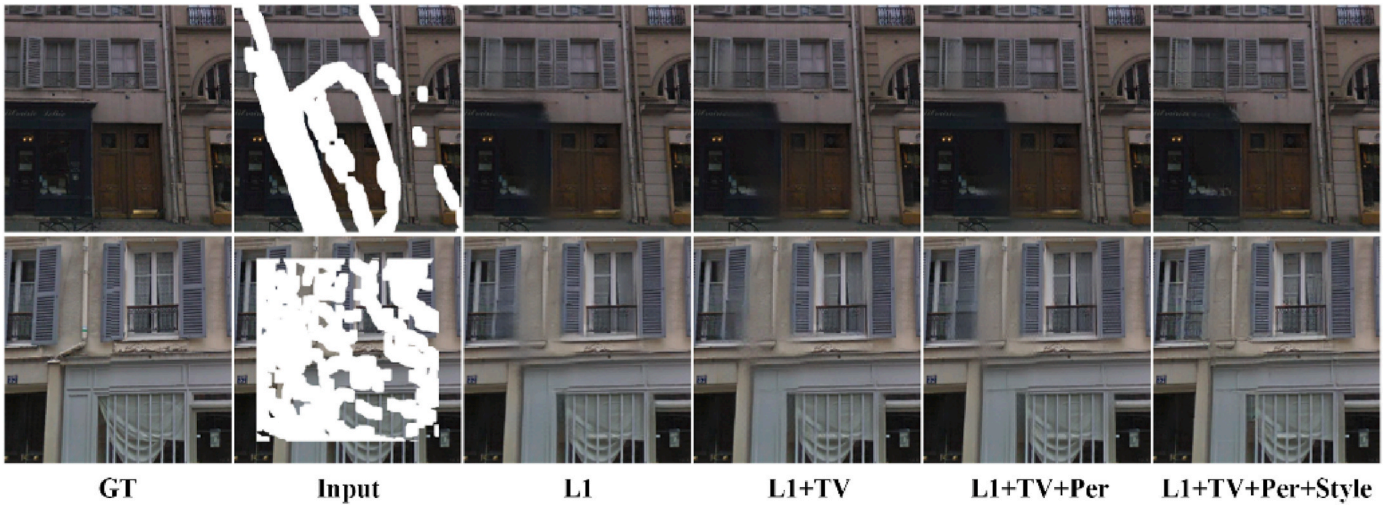


Fig. 14. Results of qualitative analysis of the loss function. Per represents perceptual loss. Zoom in for more details.

Table 9
Model computational complexity statistics.

DRFNet Stage	#Parameters	FLOPs	Infer. time/per image	Train. time/per image
Block1	94.32 M	27.07 G	8.23 ms	–
Block2	41.83 M	18.09 G	1.32 ms	–
Block3	29.25 M	17.73 G	1.11 ms	–
Block4	16.66 M	16.32 G	0.93 ms	–
Block5	1.45 M	8.00 G	0.53 ms	–
Total	183.51 M	87.21 G	12.12 ms	26.11 ms

Transconv module can effectively reduce DRFNet's parameters and decrease testing time.

3.5. Real-world applications

This subsection mainly explores application potential of DRFNet in real-world scenarios such as object removal, logo removal and editing. As shown in Fig. 15, first column and third row show ability to remove billboards and target objects from scene. The second row demonstrates certain face editing ability. In fourth row, DRFNet successfully removed the logo on truck and changed it to EAAI, which shows that DRFNet has certain ability in logo removal and re-editing.

4. Conclusion

This paper proposed a novel degression receptive field network (DRFNet) for image inpainting. DRFNet is based on gradual application of multi-scale receptive fields and mainly consists of five blocks with decreasing receptive fields, which addresses the degradation problem in multi-stage networks. In addition, to overcome the problem of local-pixels influence in pure convolutional networks, a TransConv module is designed that could be used directly to replace convolution. Extensive experiments demonstrated that the proposed model has better inpainting effects on characters, streets, and scenes. The inpainting results have fewer artifacts and reasonable semantic structure, and the qualitative and quantitative evaluations have reached state-of-the-art. However, as shown in Fig. 16, DRFNet still has limitations in constructing images with precise semantics and rich textures when inpainting large-scale damage images in complex scenes. The DRFNet model proposes a new

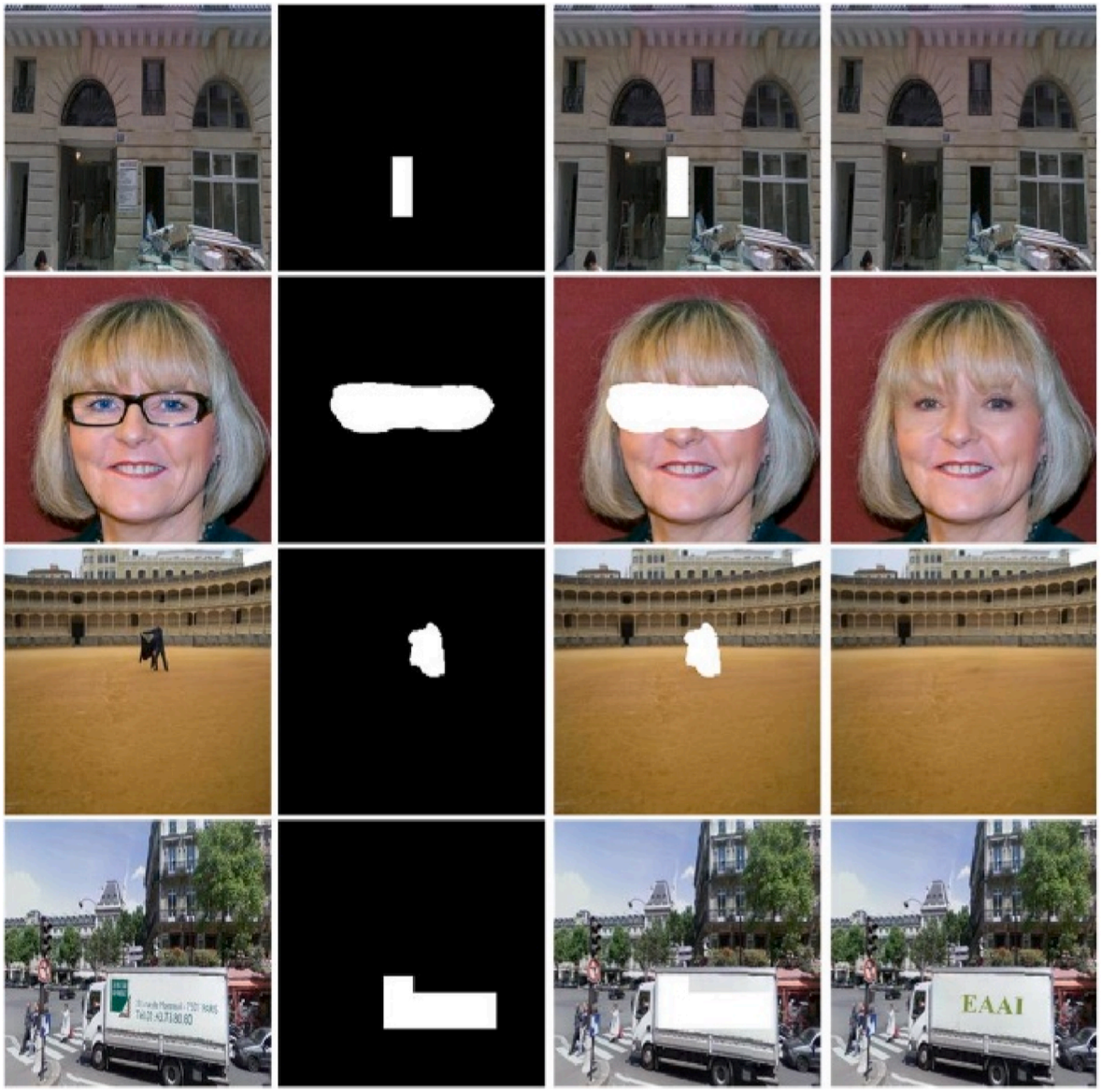


Fig. 15. Experiments for real-world applications. The first column is original image, second column is mask, third column is image to be repaired with the object to be removed, and fourth column is repaired image after removal. Zoom in for more details.

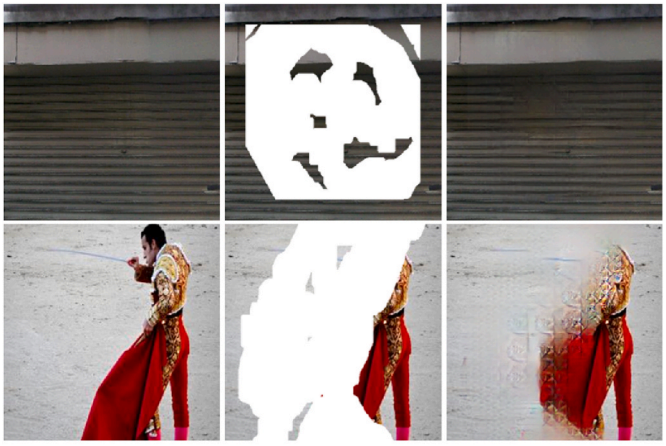


Fig. 16. Example of failure case.

design idea for inpainting networks without large increase parameters, which has enlightening significance for the design of inpainting network architectures.

In future work, incorporating semantic information to guide repairing damaged images in complex scenes will be crucial to improving inpainting performance.

CRediT authorship contribution statement

Jiahao Meng: Writing – original draft, Software, Methodology. **Weirong Liu:** Writing – review & editing, Supervision, Conceptualization. **Changhong Shi:** Validation, Formal analysis. **Zhijun Li:** Visualization, Data curation. **Chaorong Liu:** Writing – review & editing, Resources.

Declaration of competing 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.

Data availability

I have shared my code and data link on Github at https://github.com/IPCSRG/DRFNet-Inpainting.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62261032 and the Key Talent Project of Gansu Province.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.engappai.2024.109397.

References

Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B., 2009. PatchMatch: a randomized correspondence algorithm for structural image editing. ACM Trans. Graph. 28, 24.
Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C., 2000. Image inpainting. In: Proceedings of the SIGGRAPH Conference, pp. 417–424.
Bertalmio, M., Vese, L., Sapiro, G., Osher, S., 2003. Simultaneous structure and texture image inpainting. IEEE Trans. Image Process. 12, 882–889.
Cai, W., Xu, X., Xu, J., Zhang, H., Yang, H., Zhang, K., He, S., 2024. Hierarchical damage correlations for old photo restoration. Inf. Fusion 107, 102340.
Chen, G., Zhang, G., Yang, Z., Liu, W., 2023a. Multi-scale patch-GAN with edge detection for image inpainting. Appl. Intell. 53, 3917–3932.
Chen, M., Zang, S., Ai, Z., Chi, J., Yang, G., Chen, C., Yu, T., 2023b. RFA-Net: residual feature attention network for fine-grained image inpainting. Eng. Appl. Artif. Intell. 119, 105814.
Chen, Y., Xia, R., Yang, K., Zou, K., 2024a. DNNAM: image inpainting algorithm via deep neural networks and attention mechanism. Appl. Soft Comput. 154, 111392.
Chen, Y., Xia, R., Yang, K., Zou, K., 2024b. MICU: image super-resolution via multi-level information compensation and U-net. Expert Syst. Appl. 245, 123111.
Dere, V.V., Shinde, A., Vast, P., 2024. Conditional reiterative High-Fidelity GAN inversion for image editing. Pattern Recogn. 147, 110068.
Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.A., 2015. What makes Paris look like Paris? Commun. ACM 58, 103–110.
Gao, T., Wen, Y., Zhang, J., Chen, T., 2024. A novel dual-stage progressive enhancement network for single image deraining. Eng. Appl. Artif. Intell. 128, 107411.
Gatys, L.A., Ecker, A.S., Bethge, M., 2016. Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Las Vegas, USA, pp. 2414–2423.
Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing. MIT Press, Montreal, Canada, pp. 2672–2680.
Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of the International Conference on Neural Information Processing Systems. Springer, Long Beach, USA, pp. 6629–6640.
Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313, 504–507.
Iskakov, K., 2021. QD-IMD: Quick Draw Irregular Mask Dataset.
Jain, J., Zhou, Y., Yu, N., Shi, H., 2023. Keys to better image inpainting: structure and texture go hand in hand. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 208–217.
Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of the European Conference on Computer Vision. Springer, Amsterdam, The Netherlands, pp. 694–711.
Kumar, P., Gupta, V., Grover, M., 2024. Dual attention and channel transformer based generative adversarial network for restoration of the damaged artwork. Eng. Appl. Artif. Intell. 128, 107457.
Li, G., Zhang, K., Su, Y., Wang, J., 2023. Feature pre-inpainting enhanced transformer for video inpainting. Eng. Appl. Artif. Intell. 123, 106323.
Li, W., Lin, Z., Zhou, K., Qi, L., Wang, Y., Jia, J., 2022. Mat: mask-aware transformer for large hole image inpainting. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10758–10768.
Liu, G., Reda, F.A., Shih, K.J., Wang, T.-C., Tao, A., Catanzaro, B., 2018. Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European Conference on Computer Vision. Springer, Munich, Germany, pp. 85–100.
Liu, H., Jiang, B., Song, Y., Huang, W., Yang, C., 2020. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In: Proceedings of the European Conference on Computer Vision. Springer, pp. 725–741.
Liu, Z., Luo, P., Wang, X., Tang, X., 2015. Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision. IEEE, Santiago, Chile, pp. 3730–3738.
Nazeri, K., Ng, E., Joseph, T., Qureshi, F., Ebrahimi, M., 2019. EdgeConnect: structure guided image inpainting using edge prediction. In: Proceedings of the IEEE International Conference on Computer Vision Workshop. IEEE, Seoul, South Korea, pp. 3265–3274.
Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A., 2016. Context encoders: feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Las Vegas, USA, pp. 2536–2544.
Phutke, S.S., Murala, S., 2023. Image inpainting via spatial projections. Pattern Recogn. 133, 109040.
Quan, W., Zhang, R., Zhang, Y., Li, Z., Wang, J., Yan, D.-M., 2022. Image inpainting with local and global refinement. IEEE Trans. Image Process. 31, 2405–2420.
Shen, J., Chan, T.F., 2002. Mathematical models for local nontexture inpaintings. SIAM J. Appl. Math. 62, 1019–1043.
Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13, 600–612.
Xu, L., Zeng, X., Li, W., Huang, Z., 2020. Multi-granularity generative adversarial nets with reconstructive sampling for image inpainting. Neurocomputing 402, 220–234.
Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S., 2018. Generative image inpainting with contextual attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Salt Lake City, Utah USA, pp. 5505–5514.
Zeng, Y., Fu, J., Chao, H., Guo, B., 2023. Aggregated contextual transformations for high-resolution image inpainting. IEEE Trans. Visual. Comput. Graph. 29, 3266–3280.
Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Salt Lake City, Utah USA, pp. 586–595.
Zhang, R., Quan, W., Zhang, Y., Wang, J., Yan, D., 2023a. W-net: structure and texture interaction for image inpainting. IEEE Trans. Multimed. 25, 7299–7310.
Zhang, X., Zhai, D., Li, T., Zhou, Y., Lin, Y., 2023b. Image inpainting based on deep learning: a review. Inf. Fusion 90, 74–94.
Zhang, Y., Yang, M., Xiao, T., Wang, Z., Chi, Z., 2024. Freezing partial source representations matters for image inpainting under limited data. Eng. Appl. Artif. Intell. 133, 108072.
Zheng, C., Cham, T.-J., Cai, J., Phung, D., 2022. Bridging global context interactions for high-fidelity image completion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, New Orleans, Louisiana USA, pp. 11512–11522.
Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A., 2017. Places: a 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. 40, 1452–1464.
Zhu, M., He, D., Li, X., Li, C., Li, F., Liu, X., Ding, E., Zhang, Z., 2021. Image inpainting by end-to-end cascaded refinement with mask awareness. IEEE Trans. Image Process. 30, 4855–4866.
Zuo, Z., Zhao, L., Li, A., Wang, Z., Zhang, Z., Chen, J., Xing, W., Lu, D., 2023. Generative image inpainting with segmentation confusion adversarial training and contrastive learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 3888–3896.

Engineering Applications of Artificial Intelligence 159 (2025) 111769

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Self-information and prediction mask enhanced blind inpainting network for dunhuang murals

Jiahao Meng, Weirong Liu, Changhong Shi, Zhijun Li, Jie Liu *

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

ARTICLE INFO

Keywords:
Blind image inpainting
Dunhuang mural images
Generative adversarial network
Information enhanced transformer block

ABSTRACT

Blind image inpainting methods based on deep learning have shown promising results in digital image inpainting of dunhuang mural images in recent years. However, current blind inpainting methods still suffer from color patches and structural confusion in the repair results caused by contamination of damaged features and sub-network interference. To address the above problems, a self-information and prediction mask enhanced blind inpainting network (SIME-BINet) for dunhuang mural images is proposed. SIME-BINet redesigns blind inpainting method of phased guidance paradigm into information enhance paradigm, which continuously optimizes enhanced information in dynamic form during training process and provides guidance for encoding process. Meanwhile, an information enhanced transformer block is designed to overcome the problem of damaged feature contamination by introducing enhanced information. Experiments show that SIME-BINet outperforms recent state-of-the-art blind inpainting methods on DhMurals1714 dataset and real damage mask. SIME-BINet offers a new paradigm for blind image inpainting based on deep learning and provides an innovative approach for inpainting of dunhuang mural images. The code, data, and pre-trained models will be made available at https://github.com/IPCSRG/SIME-BINet after the paper is published.

1. Introduction

The dunhuang murals are not only treasures of ancient Chinese art but also a significant part of the world cultural heritage (Hao and Qiaomei, 2010). They demonstrate cultural exchange and integration between ancient China and various national cultures around the world. However, the dunhuang murals have suffered damage due to environmental conditions, natural disasters, and human activities. These damages not only affect the preservation and display of the murals but also cause immeasurable losses to their artistic and historical value.

The restoration of dunhuang murals began as early as the 1940s. At that time, mural restoration mainly involved direct manual overpainting on the original walls, which often carried a high risk of secondary damage to the murals. In the 1990s, digital inpainting technology was gradually introduced into mural restoration. This technology helps to avoid the irreversible shortcomings of direct repair and allows more room for trial and error.

Early digital image inpainting methods mainly include damaged edge diffusion methods (Bertalmio et al., 2003; Chan and Shen, 2001; Shen and Chan, 2002) and sample match methods (Barnes et al., 2009; Criminisi et al., 2004; Le Meur et al., 2011; Wang et al., 2017). The damaged edge diffusion method is mainly based on valid pixels around the damaged area and gradually propagates to the undamaged area. This approach is unable to repair structure and texture when large areas are damaged, although some progress has been made in repairing small damage. Sample match methods mainly use similar information from the undamaged area to fill the damaged area, and such methods have been successful in inpainting simple mural images with strong symmetry. However, sample match methods are unable to repair mural images with relatively complex content.

In recent years, with the advancement of technology, image inpainting has been widely applied in various fields (Gao et al., 2024b; Mosleh et al., 2013, 2017; Peng and Chellappa, 2023; Zhang et al., 2023) such as object removal, cultural relic restoration, and image deblurring. The inpainting of digital Dunhuang murals has also increasingly turned to the field of deep learning. Based on the damage mask requirement, deep learning-based image inpainting methods can be categorized into non-blind inpainting methods and blind inpainting methods (Wang

* Corresponding author.
E-mail addresses: mjhforwork@163.com (J. Meng), liuwr@lut.edu.cn (W. Liu), changhong_shi@126.com (C. Shi), lizhijun_lut@163.com (Z. Li), ljdaisy@163.com (J. Liu).

https://doi.org/10.1016/j.engappai.2025.111769
Received 8 February 2025; Received in revised form 14 June 2025; Accepted 10 July 2025
Available online 17 July 2025

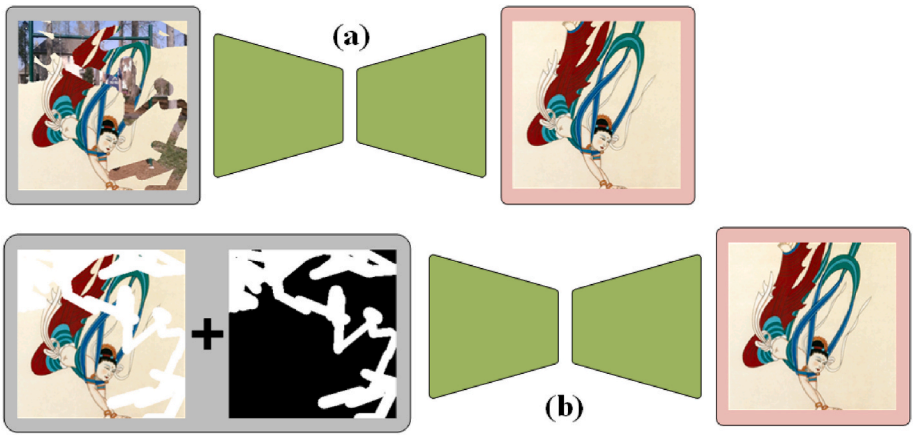


Fig. 1. Image inpainting methods based on deep learning. (a) Blind inpainting methods. (b) Non-blind inpainting methods.

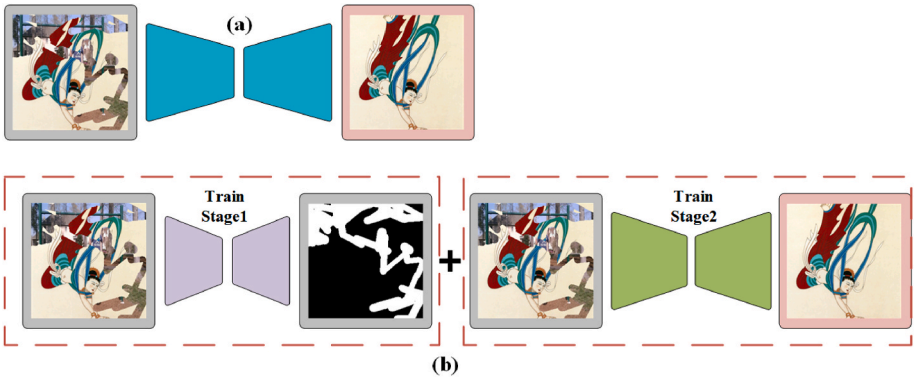


Fig. 2. The type of blind inpainting methods. (a) Direct blind inpainting methods. (b) Blind inpainting methods guided by mask prediction.

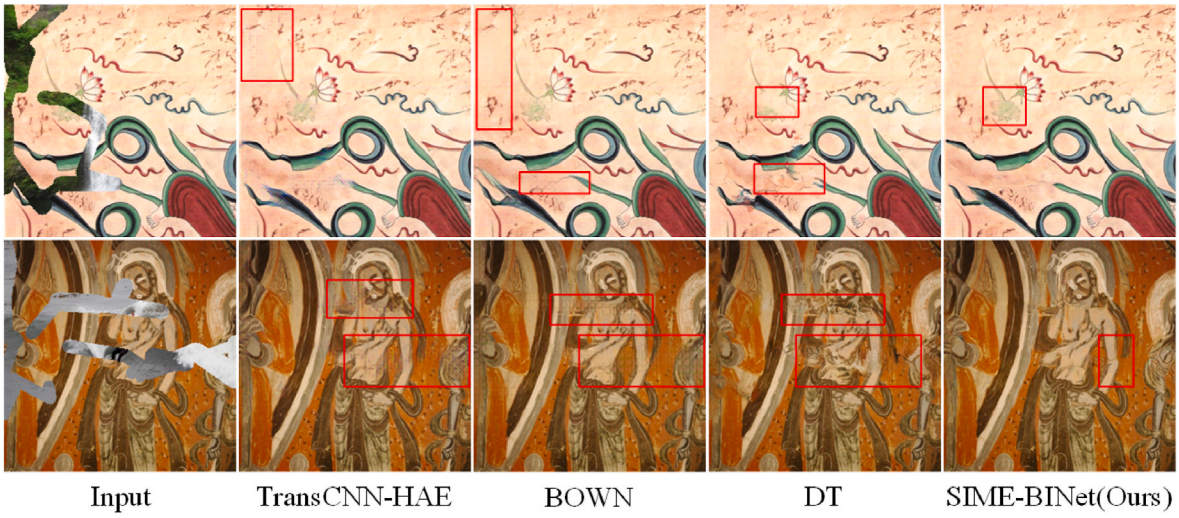


Fig. 3. Example of experimental phenomenon. Please download the supplementary materials for more details.

et al., 2023). Non-blind inpainting methods (Chen et al., 2023, 2024a, 2024b, 2024c; Deng and Yu, 2023; Huang and Huang, 2023; Iizuka et al., 2017; Liu et al., 2023; Meng et al., 2024; Pathak et al., 2016; Ren et al., 2024; Yu et al., 2018; Zhou et al., 2022) assume that the damaged region is known, and the image is inpainted according to undamaged area, as shown in Fig. 1 (a). However, non-blind inpainting methods cannot work properly without masking information, since the mask of damaged region is not always provided. Therefore, some researchers have increasingly focused on developing blind inpainting methods (Kumar et al., 2024; Li et al., 2023, 2024; Phutke et al., 2023; Schmidt et al., 2022; Wang et al., 2020; Zhao et al., 2022, 2024a, 2024b) that do not require damage masks, as shown in Fig. 1 (b).

Current blind inpainting methods can be divided into two approaches: direct blind inpainting (DBI) methods, as shown in Fig. 2 (a), and mask prediction guided blind inpainting (MGBI) methods, as shown in Fig. 2 (b). Experimental analysis of current open-source blind

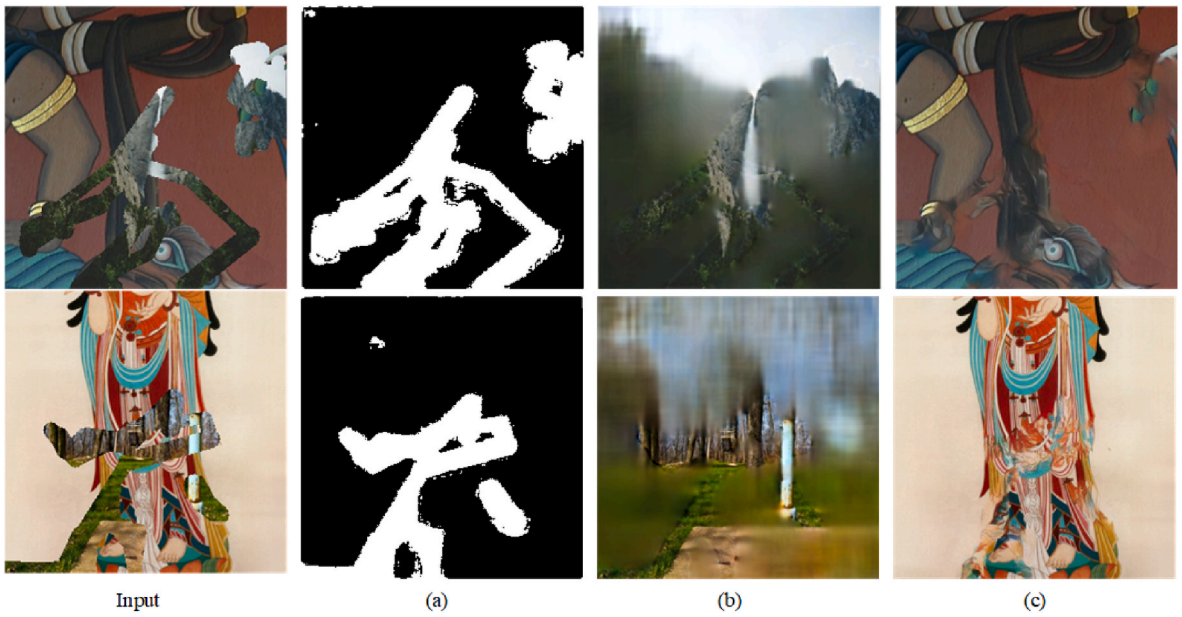


Fig. 4. Analysis of the DT. (a) represents the binary result of MPN prediction mask. (b) represents the prediction result of CPN. (c) Shows the repair result of DT.

inpainting methods reveals certain limitations. In particular, transformer-based DBI methods suffer from feature contamination problem because they have difficulty distinguishing between features of damaged and undamaged regions. The problem of feature contamination easily leads to color patches, such as TransCNN-HAE model and BOWN model in Fig. 3.

On the other hand, MGBI methods are typically multi-stage training frameworks that rely on the intermediate result of previous sub-network, and any inaccuracies in predictions at earlier stages can easily propagate. The problems of sub-network interference in this approach not only cause structural confusion in repaired areas by introducing unreliable prediction information into the network but also increase the complexity of training. For example, the blind inpainting method Decontamination Transformer (DT) is mainly composed of three parts: mask prediction network (MPN), contaminant prediction network (CPN), and image inpainting network (IIN). In the training phase, the MPN and CPN&INN networks need to be trained separately. Then, the three sub-networks MPN, CPN, and INN need to be combined for fine-tuning. For testing, the MPN prediction mask is used to obtain the noise prediction mapping features, and the noise mapping features are used to guide the repair of the damaged image. In particular, MPN uses soft-mask prediction, which can easily lead to the misalignment of noisy mapping features. In order to facilitate the analysis and comparison, the data greater than 0 in the MPN network prediction mask are reset to 1 to form a binary mask, as shown in Fig. 4 (a). The soft-mask causes discontinuous areas at the edges of the prediction mask and residual pixel values within the damaged region. This leads to inaccurate prediction noise in the CPN and affects the repair results, as shown in Fig. 4 (c).

Inspired by the above analysis, an end-to-end self-information and prediction mask enhanced blind inpainting network for dunhuang mural images is proposed, which addresses the problems of color patches and structural confusion in current blind inpainting methods. Firstly, a transformer-based encoder is designed to leverage the ability of transformers to map long-range dependencies. Secondly, an enhanced information generation module (EIGM) is proposed to redesign the phased guidance paradigm of the mask prediction-guided blind inpainting method into an information enhancement paradigm. This paradigm dynamically optimizes the enhanced information during the training process and guides the encoding process. Finally, an information enhanced transformer block (IETSBlock) is designed to constrain the training process by introducing enhanced information features from EIGM. The main contributions are as follows.

1. A novel end-to-end self-information and prediction mask enhanced blind inpainting network (SIME-BINet) for dunhuang mural images is proposed to address the problem of sub-network interference. SIME-BINet has fewer color patches and achieves better inpainting consistency compared to the recent state-of-the-art blind inpainting models.
2. An enhanced information generation module (EIGM) is proposed to continuously optimize enhanced information in dynamic manner. The EIGM reduces the problem of incorrect prediction of intermediate features caused by soft-mask in current blind inpainting methods.
3. An information enhanced transformer block (IETSBlock) is proposed to address the problem of feature contamination in blind inpainting models by introducing enhanced information to constrain multi-head attention.

2. Related work

2.1. Non-blind image inpainting

In 2016, Pathak et al. (2016) made a breakthrough in image inpainting through the application of deep learning and named context-encoder(CE). Many subsequent works have been gradually developed based on CE model. Iizuka et al. (2017) proposed a globally and locally (GL) image inpainting method that uses global and local discriminators for training to ensure consistency of global and local semantic information, respectively. However, CE-based image inpainting models cannot adapt to different types of damage. Yu et al. (2018) proposed contextual attention (CA) mechanism based on GL model and further designed a coarse-to-fine two-stage image inpainting network, which can repair any type of damaged image. But, CA model has visible repair traces when repairing large irregular damage. Zhou et al. (2022) proposed a structure-guided image inpainting method for dunhuang mural image inpainting, which uses relevant color information in deep features to improve the color quality of structural regions. This method has limitations on inpainting damage of different forms and larger sizes. Chen et al. (2023) proposed multi-scale patch-gan combined with edge detection for image inpainting. However, this method only shows the

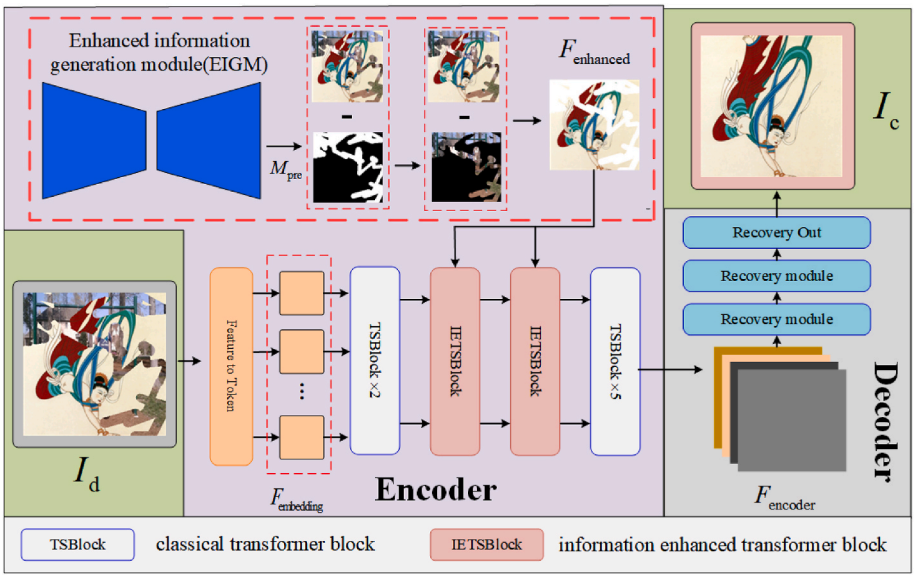


Fig. 5. The architecture of proposed blind inpainting network for dunhuang mural images.

repair results for center damage, and the performance for other types of damage is unknown. Liu et al. (2023) proposed a multi-stage progressive reasoning network with global and local receptive fields for dunhuang mural image inpainting, which can recursively infer structural and texture features from hole boundaries and gradually tighten constraints on hole centers. Due to the progressive reasoning structure, the network has more redundant information. Deng et al. (Deng and Yu, 2023) proposed a structure-guided two-branch (SGTB) model for repairing damaged areas of ancient murals. The generalization performance of SGTB is limited by the lack of high-quality data. Ren et al. (2024) proposed a dunhuang mural image inpainting model that combines a parallel double convolutional feature extraction depth generator and a ternary heterogeneous joint discriminator. This method has problems with color distortion or brightness differences. Chen et al. (2024b) proposed a two-stage image inpainting network based on the inference attention module. However, the repair results of the network are susceptible to visual bias when the damaged area increases.

Although non-blind image inpainting frameworks have demonstrated advancements in content coherence and structural plausibility, their dependence on damaged region annotations restricts applicability in real-world scenarios. The dependence on precise damage masks has motivated recent research in blind image inpainting based on deep learning.

2.2. Blind image inpainting

Wang et al. (2020) proposed a two-stage visual coherence network including a mask prediction Network (MPN) and a robust inpainting network (RIN) for blind image inpainting. This model has the problem of being unable to distinguish between undamaged areas when the damaged areas are large. Schmidt et al. (2022) combined CAR (Sun and Chen, 2020) and HINet (Chen et al., 2021) models and proposed an adaptive resampled instance normalized blind inpainting network that can repair damaged dunhuang mural images without using any prior degradation information. This network has the problem of losing reconstruction details. Zhao et al. (2022) proposed a single-stage model for blind image inpainting by using the global long-distance context modeling capability of the transformer and the local short-distance context modeling ability of CNN. Because the damaged regions are not distinguished, the transformer module in the network has a feature contamination problem. Li et al. (2023) proposed a blind image inpainting model composed of mask prediction network (MPN), contaminant prediction network (CPN), and image inpainting network (IIN). However, the network has the problem of inaccurate prediction caused by soft-mask. Phutke et al. (2023) introduced an end-to-end transformer architecture for blind image inpainting that improves repair performance through full-dimensional gated attention and wavelet queries. Zhao et al. (2024a) proposed a heterogeneous blind inpainting method based on transformer and u-net for dunhuang mural images. These transformer-based networks also suffers from feature contamination due to the inability to distinguish between damaged and undamaged regions. Li et al. (2024) proposed a two-stage network to locate semantically inconsistent regions through global semantic features and generate reasonable content. The network has problems in distinguishing the contaminated area from the image background. Kumar et al. (2024) proposed a new generative adversarial network based on dual-attention mechanism for damaged artwork image inpainting. However, the limitation of this method is that the original version of each damaged artwork is required.

The above blind inpainting methods liberate deep learning-based image inpainting task from masks. However, current blind inpainting methods still suffer from color patches and structural confusion in the repair results caused by contamination of damaged features and sub-network interference.

3. Method

3.1. Overall architecture

A self-information and prediction mask enhanced blind inpainting network (SIME-BINet) for dunhuang mural images is proposed to address the problems of feature contamination and sub-network interference in blind inpainting methods. As shown in Fig. 5, SIME-BINet is an end-to-end blind inpainting network consist of two parts, encoder with enhanced information generation module and decoder.

For blind inpainting task, assume that model input contains only damaged dunhuang mural images $\boldsymbol{I}_{\mathrm{d}} \in \mathbb{R}^{\boldsymbol{c}\times\boldsymbol{h}\times\boldsymbol{w}}$ and visually complete dunhuang mural images $\boldsymbol{I}_{\mathrm{c}} \in \mathbb{R}^{\boldsymbol{c}\times\boldsymbol{h}\times\boldsymbol{w}}$ is expected to be obtained directly by accurately remove contaminated area from $\boldsymbol{I}_{\mathrm{d}}$.

Firstly, taking damaged image $\boldsymbol{I}_{\mathrm{d}}$ as input, the enhancement information generation module (EIGM) is designed to reduce the incorrect prediction of intermediate features, which can be defined as follows.

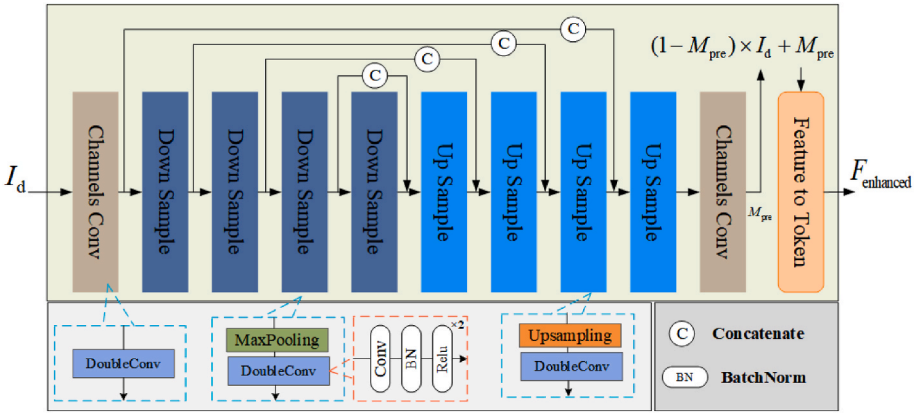


Fig. 6. Structure of enhanced information generation module.

$$\begin{aligned} \boldsymbol{F}_{\text{enhanced}} &= \text{EIGM}(\boldsymbol{I}_{\text{d}}) \\ &= \text{ChannelConv}(\text{DownSample}(\text{Upsample}(\text{ChannelConv}(\text{FtoT}(\boldsymbol{I}_{\text{d}}))))) \end{aligned} \tag{1}$$

where $\text{EIGM}(\cdot)$ denotes function of enhancement information generation module, $\text{FtoT}(\cdot)$ represents converting features into tokens, and $\boldsymbol{F}_{\text{enhanced}} \in \mathbb{R}^{t\times \boldsymbol{e}}$ represents feature map generated by $\text{EIGM}(\cdot)$ with mask information and valid self-information.

Then, transformer-based encoder is designed to extract features from murals image and establish long-range dependencies within the image, which is defined as follows.

$$\begin{aligned} \boldsymbol{F}_{\text{encoder}} &= \text{Encoder}\left(\text{FtoT}(\boldsymbol{I}_{\text{d}}), \boldsymbol{f}_{\text{EIGM}}(\boldsymbol{I}_{\text{d}})\right) = \boldsymbol{f}_{\text{encoder}}\left(\boldsymbol{F}_{\text{embedding}}, \boldsymbol{F}_{\text{enhanced}}\right) \\ &= \text{TSBlock}\left(\text{IETSBlock}\left(\text{TSBlock}\left(\boldsymbol{F}_{\text{embedding}}\right), \boldsymbol{F}_{\text{enhanced}}\right)\right) \end{aligned} \tag{2}$$

where $\text{Encoder}(\cdot)$ represents function of encoder and $\boldsymbol{F}_{\text{encoder}}$ is feature map from encoder output. $\text{TSBlock}(\cdot)$ represents classical transformer block and $\text{IETSBlock}(\cdot)$ represents information enhanced transformer block, as shown in Fig. 7.

Finally, decoder is constructed to recover murals image from encoded feature $\boldsymbol{F}_{\text{encoder}}$, and can be defined as follows.

$$\boldsymbol{I}_{\text{c}} = \text{Decoder}(\boldsymbol{F}_{\text{encoder}}) = \text{RecoveryOut}(\text{RecoveryModule}(\text{FtoT}(\boldsymbol{F}_{\text{encoder}}))) \tag{3}$$

where $\text{Decoder}(\cdot)$ represents function of decoder. The structure of $\text{RecoveryModule}(\cdot)$ and $\text{RecoveryOut}(\cdot)$ is shown in Fig. 8. $\boldsymbol{I}_{\text{c}}$ is visually complete dunhuang mural image that is expected to be obtained. The overall inference process is shown in Algorithm 1.

Algorithm 1. Inference pseudo-code of SIME-BINet.

Input: damaged dunhuang mural images $\boldsymbol{I}_{\text{d}} \in \mathbb{R}^{\boldsymbol{c}\times \boldsymbol{h}\times \boldsymbol{w}}$;
Output: visually complete dunhuang mural images $\boldsymbol{I}_{\text{c}} \in \mathbb{R}^{\boldsymbol{c}\times \boldsymbol{h}\times \boldsymbol{w}}$;
Inference:
  for *batchsize*←1 **to** *total_test_sample* **do**
    $\boldsymbol{F}_{\text{enhanced}} \leftarrow \text{EIGM}(\boldsymbol{I}_{\text{d}})$//enhancement information
    $\boldsymbol{F}_{\text{embedding}} \leftarrow \text{FtoT}(\boldsymbol{I}_{\text{d}})$//converting features into tokens
    $\boldsymbol{F}_{\text{encoder}} \leftarrow \text{Encoder}\left(\left(\boldsymbol{F}_{\text{embedding}}, \boldsymbol{F}_{\text{enhanced}}\right)\right)$//the enhanced information is used for encoding
    $\boldsymbol{I}_{\text{c}} \leftarrow \text{Decoder}(\boldsymbol{F}_{\text{encoder}})$//visually complete dunhuang mural image
    save $\boldsymbol{I}_{\text{c}}$ to image file
  end for

3.2. Encoder

To solve the sub-network interference problem in the multi-stage training blind inpainting method. SIME-BINet introduces enhancement information generation module (EIGM) into encoder, which uses mask prediction and valid regions as enhancement information instead of guidance, and then designs information enhancement transformer block.

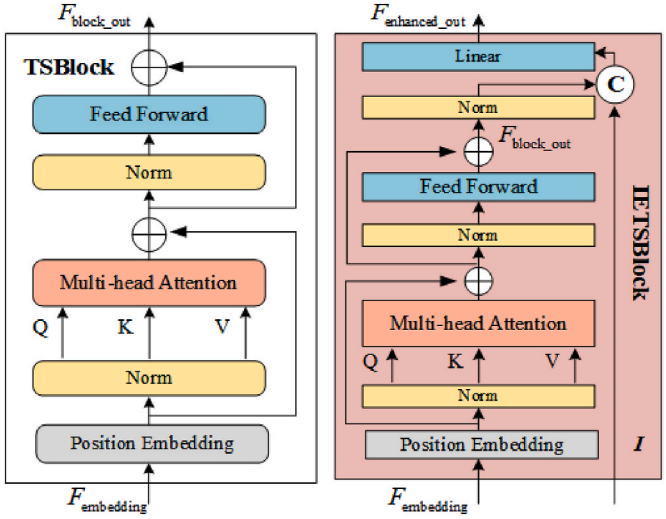


Fig. 7. Comparison of information enhanced transformer block and traditional transformer block.

As shown in Fig. 5, encoder mainly consists of transformer block (TSBlock) and information enhanced transformer block (IETSBlock). Firstly, damaged image $\boldsymbol{I}_{\text{d}}$ as input, $\boldsymbol{I}_{\text{d}}$ is sliced into tokens with contextual features for transformer block operation, and mapped to $\boldsymbol{F}_{\text{embedding}} \in \mathbb{R}^{t\times \boldsymbol{e}}$ by patch embedding, where e denotes embedded dimension(Gao et al., 2024a), $t = \frac{h}{p} \times \frac{w}{p}$, p denotes patch size, h and w are height and width, respectively. Then, TSBlock is used to reconstruct and encode relationship between features, which in turn introduces enhancement information through IETSBlock for continuous optimization of encoder process. Finally, optimized features are further reconstructed by TSBlock. The encoder process is shown as follows:

$$\boldsymbol{F}_{\text{encoder}} = \text{TSBlock}\left(\text{IETSBlock}\left(\text{TSBlock}\left(\boldsymbol{F}_{\text{embedding}}\right), \boldsymbol{F}_{\text{enhanced}}\right)\right) \tag{4}$$

where $\text{TSBlock}(\cdot)$ represents classical transformer block and $\text{IETSBlock}(\cdot)$ represents information enhanced transformer block.

3.2.1. Enhanced information generation module

Enhancement information generation module (EIGM) is designed to continuously optimize enhancement information during training process in dynamic form and provide guidance for encoder process, with structure shown in Fig. 6.

Firstly, $\boldsymbol{I}_{\text{d}}$ as input, and input feature dimension was extended

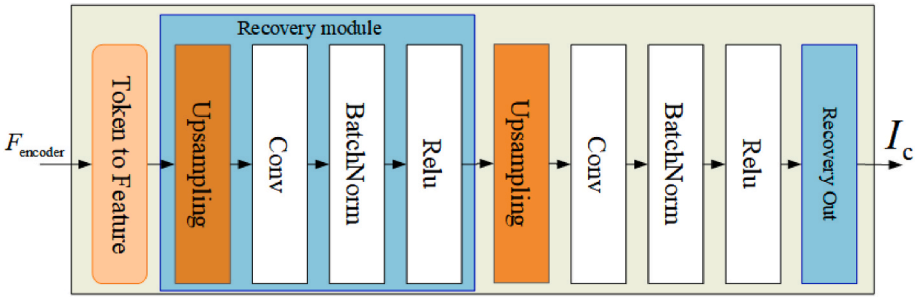


Fig. 8. Schematic of decoder structure.

through channels conv to ensure diversity of features. Secondly, four groups of down samples were used to gradually downsample features to obtain high-dimension feature group. Thirdly, four groups of up samples were used to gradually recover low-dimension features. Then feature dimensions are recovered by channels conv to get predicted damaged region mask $\boldsymbol{M}_{\text{pre}}$. Finally, $\boldsymbol{F}_{\text{enhanced}}$ is obtained from $\boldsymbol{M}_{\text{pre}}$ and $\boldsymbol{I}_{\text{d}}$, where $\boldsymbol{F}_{\text{enhanced}}$ only consists of damaged mask and undamaged features.

3.2.2. Information enhanced transformer block

Information Enhanced transformer block (IETSBlock) is constructed to address feature contamination problem of transformer block. Specifically, by introduce enhanced information $\boldsymbol{F}_{\text{enhanced}}$ only contain damaged mask and undamaged regions features to constraints training process. Compared with TSBlock as shown in Fig. 7, IETSBlock introduces enhancement information $\boldsymbol{I}$ and concatenates $\boldsymbol{I}$ with original output features $\boldsymbol{F}_{\text{block\_out}}$, followed by fusion and transformation of features by linear regression. The details are as follows:

$$\boldsymbol{F}_{\text{enhanced\_out}} = \text{Linear}(\text{Cat}(\text{Norm}(\boldsymbol{F}_{\text{block\_out}}), \boldsymbol{I})) \tag{5}$$

where $\text{Norm}(\cdot)$ denotes normalization operation, $\text{Cat}(\cdot)$ represents Concatenate operation, $\text{Linear}(\cdot)$ linear regression operation, and $\boldsymbol{I}$ is $\boldsymbol{F}_{\text{enhanced}}$ in SIME-BINet.

3.3. Decoder

Inpainted image $\boldsymbol{I}_{\text{c}}$ is obtained by recover encoded high-dimension features $\boldsymbol{F}_{\text{encoder}}$ into detailed low-dimension features. Firstly, decoder converts token $\boldsymbol{F}_{\text{encoder}} \in \mathbb{R}^{\boldsymbol{t}\times\boldsymbol{e}}$ to map $\boldsymbol{F}_{\text{encoder\_f}} \in \mathbb{R}^{\boldsymbol{c}'\times\boldsymbol{h}'\times\boldsymbol{w}'}$, where $h' = w' = \frac{h}{p} = \frac{w}{p}$, $c' = h = w$. Then, two recovery module are used to refine $\boldsymbol{F}_{\text{encoder\_f}}$, as shown in follow.

$$\boldsymbol{F}_{\text{recovery}} = \text{Relu}\left(\text{BatchNorm}\left(\text{Conv}\left(\text{Upsampling}\left(\boldsymbol{F}_{\text{encoder\_f}}\right)\right)\right)\right) \tag{6}$$

where $\boldsymbol{F}_{\text{recovery}} \in \mathbb{R}^{\boldsymbol{c}''\times\boldsymbol{h}\times\boldsymbol{w}}$ denotes feature that has been refined by recovery module, $c'' = h' = w'$.

Finally, feature $\boldsymbol{F}_{\text{recovery}}$ is restored to inpainted image $\boldsymbol{I}_{\text{c}} \in \mathbb{R}^{\boldsymbol{c}\times\boldsymbol{h}\times\boldsymbol{w}}$.

$$\boldsymbol{I}_{\text{c}} = \text{RecoveryOut}\left(\boldsymbol{F}_{\text{recovery}}\right) \tag{7}$$

where $\text{RecoveryOut}(\cdot)$ indicates channel filter operation.

3.4. Loss functions

SIME-BINet introduces composite loss function for efficient training. Firstly, adversarial loss is performed to transform network optimization problem into minimax optimization problem to approximates the distribution of repaired image $\boldsymbol{I}_{\text{c}}$ to real image $\boldsymbol{I}_{\text{gt}}$.

$$\begin{aligned}\ell_{\text{Adv}} = \min_{\text{G}}\max_{\text{D}} & \mathbb{E}_{\boldsymbol{I}_{\text{gt}}\sim\mathbb{P}_{\text{r}}}\left[\boldsymbol{log}(\text{D}(\boldsymbol{I}_{\text{gt}}))\right] \\ & -\mathbb{E}_{\boldsymbol{I}_{\text{c}}\sim\mathbb{P}_{\text{g}}}\left[\boldsymbol{log}(\text{D}(\boldsymbol{I}_{\text{c}}))\right]\end{aligned} \tag{8}$$

where $\text{D}(\cdot)$ denotes the feature extraction network using the pre-trained VGG16 network for $\boldsymbol{I}_{\text{gt}}$ and $\boldsymbol{I}_{\text{c}}$, respectively (Simonyan and Zisserman).

Secondly, mask prediction loss is designed to continuously optimize accuracy of prediction mask, which is composed of BCE loss (Ruby and Yendapalli, 2020) and DICE loss (Li et al., 2020).

$$\ell_{\text{BCE}} = -\left[\boldsymbol{M}\log(\boldsymbol{M}_{\text{pre}}) + (1-\boldsymbol{M})\log(1-\boldsymbol{M}_{\text{pre}})\right] \tag{9}$$

$$\ell_{\text{DICE}} = \left(1 - \frac{2\times\left|\boldsymbol{p}_{\text{pre}}\cap\boldsymbol{p}\right|}{\left|\boldsymbol{p}_{\text{pre}}\right| + |\boldsymbol{p}|}\right) \tag{10}$$

$$\ell_{\text{M}} = \ell_{\text{BCE}} + \ell_{\text{DICE}} \tag{11}$$

where BCE loss is used to measure difference between predicted probability and true labels, DICE loss is used to measure similarity between predicted results and true labels, $\left|\boldsymbol{p}_{\text{pre}}\right|$ indicates the number of pixels in predicted results and $|\boldsymbol{p}|$ indicates the number of pixels in true labels.

Finally, L1 loss is introduced to further optimize training process to improve pixel-level reconstruction effect, and total variation (TV) loss (Liu et al., 2018) is introduced to enhance image smoothness. At the same time, perceptual loss (Johnson et al., 2016) and style loss (Gatys et al., 2016) are also introduced into loss function to improve ability of structure reconstruction and texture recovery. Generator loss function is:

$$\ell_{\text{G}} = \ell_{\text{L1}} + \lambda_1\ell_{\text{tv}} + \lambda_2\ell_{\text{perceptual}} + \lambda_3\ell_{\text{style}} \tag{12}$$

where λ_1, λ_2, λ_3 denote regularization parameters for each loss.

In summary, proposed SIME-BINet is trained in end-to-end fashion, and total loss function is as follows:

$$\ell_{\text{total}} = \ell_{\text{Adv}} + \ell_{\text{M}} + \ell_{\text{G}} \tag{13}$$

4. Experiments

4.1. Experimental settings

4.1.1. Datasets

The dunhuang mural image dataset (Li et al.) is used for experimental analysis to evaluate the effectiveness of SIME-BINet, as shown in Table 1. Standard non-blind mural image inpainting model simulates that damage masks are condition and binary masks are used to simulate damage for train and test, as shown in Fig. 9. However, in blind inpainting task of mural images, mask information is unknown in testing, and using only monochrome masks will lead to poor generalization ability of model due to diverse damage of mural images. Therefore, in order to simulate complex damage of mural images and

Table 1
Details of the dunhuang mural image dataset.

Dataset	Train data	Val data	Test data	Size
DhMurals1714	1564	100	50	256 × 256

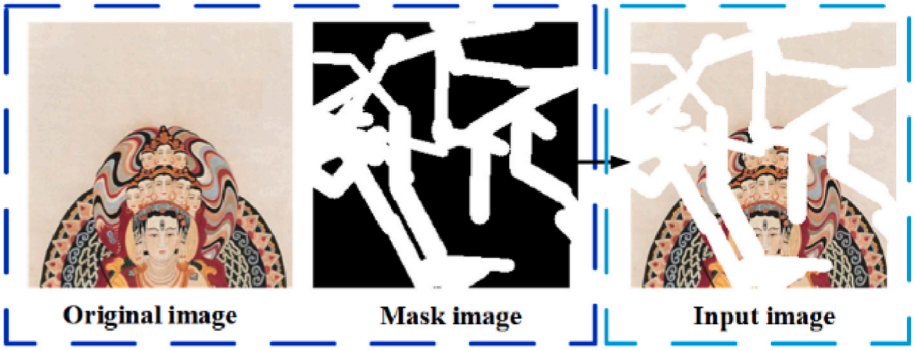


Fig. 9. Sample data from blind inpainting task of dunhuang mural images(Li et al.).

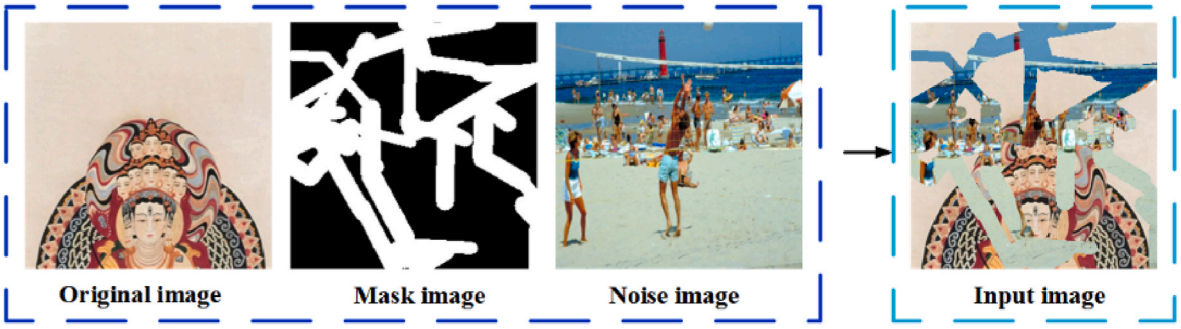


Fig. 10. Sample data from blind inpainting task of dunhuang mural images(Wang et al., 2020).

Fig. 11. Sample data from real damage masks.

improve the generalization ability and robustness of the model (Li et al., 2024). Complex natural scene dataset Places2 (Zhou et al., 2017) is introduced as simulated damaged noise fill to construct a more complex simulated damaged dataset following blind inpainting literature (Zhao et al., 2022), as shown in Fig. 10.

In addition, different from other mural image inpainting methods that test through simulated masks as shown in Figs. 9 and 10, real damage masks are constructed to further the verify effectiveness of SIME-BINet, as shown in Fig. 11.

4.1.2. Comparison methods

Three state-of-the-art (SOTA) blind inpainting models TransCNN-HAE, BOWN, and DT, over the past three years are selected for comparative experiments, as few open-source blind inpainting models are available.

- **TransCNN-HAE**: one-stage Transformer-CNN hybrid autoencoder for blind image inpainting Zhao et al. (2022).
- **BOWN**: one-stage transformer-base approach for blind image inpainting Phutke et al. (2023).
- **DT**: multi-stage decontamination transformer model for blind image inpainting Li et al. (2023).

4.1.3. Implementation details

Train batch size is set to 4 and Adam with $\beta_1 = 0.5$, $\beta_2 = 0.999$ used to optimizer network in end-to-end manner. Decay learning strategy is used where total epoch = 400, initial learning rate = 0.0002 for first 100 Epochs, and learning rate of last 300 Epochs is linearly decayed to 0. Embedded dimension e is 256,patch size $p = 4$. Regularization parameters $\lambda_1 = 0.1$, $\lambda_2 = 0.05$, and $\lambda_3 = 120$. Intel Core i7-13700 CPU and single NVIDIA GeForce RTX 4090 GPU are used for training.

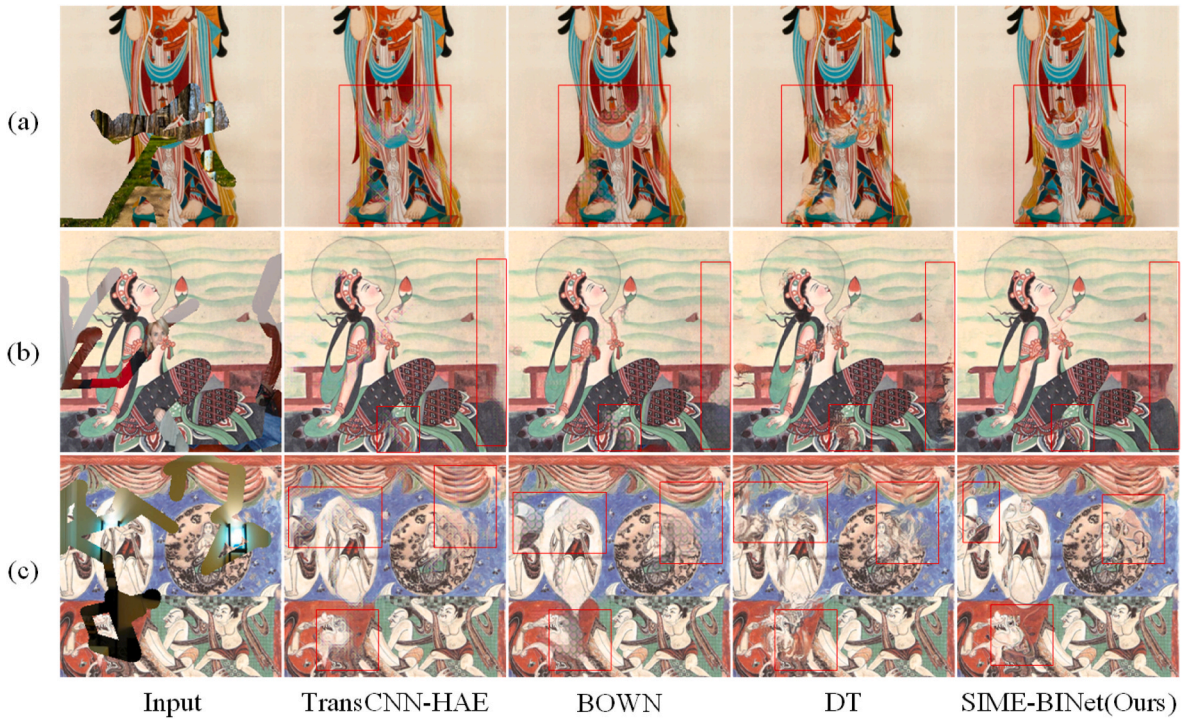


Fig. 12. Examples of qualitative comparison results between SIME-BINet and TransCNN-HAE, BOWN, and DT. Please download the supplementary materials for more details.

Table 2
Quantitative comparison results for TransCNN-HAE, BOWN, DT, and Ours. ↑ and ↓ represent larger and smaller is better, respectively. The percentages are calculated using TransCNN-HAE as the baseline. ↑ represents the percentage increase in performance compared to baseline, while ↓ represents the percentage decrease in performance.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
TransCNN-HAE	24.4894 0	0.8646 0	0.0794 0
BOWN	24.7980 ↑ 1.26 %	0.8752 ↑ 1.23 %	0.0827 ↓ 4.16 %
DT	23.0748 ↓ 5.78 %	0.8430 ↓ 2.50 %	0.1046 ↓ 31.74 %
SIME-BINet(Ours)	**25.0997** ↑ 2.49 %	**0.8837** ↑ 2.21 %	**0.0701** ↑ 11.71 %

4.2. Experimental results and analysis

Qualitative and quantitative experiments were performed to compare SIME-BINet with the recent SOTA blind inpainting method. In addition, official open source code, official configuration, same dataset, and hardware platform are used for train and test to ensure fairness of comparison experiments.

4.2.1. Qualitative comparisons

Sample results of qualitative comparison between SIME-BINet, TransCNN-HAE, BOWN, and DT are shown in Fig. 12. Repair results of SIME-BINet have no obvious artifacts and better visual results. In contrast, TransCNN-HAE and BOWN have obvious grid-like color patches due to the lack of a clear guide to damaged region. The multi-stage blind inpainting method DT has misstructured structure due to the misalignment of predictions in the intermediate stage. For example, focus on red-boxed areas in Fig. 12. In the first row, at the left foot of Buddha, SIME-BINet has clearer lines and is not affected by color stains compared to other comparative methods. In the second row, at the lower middle red area, the textural details and visual effects are better than other methods, although SIME-BINet does not repair complete peripheral details of the cushion.

Table 3
Qualitative comparison results for single dunhuang mural image in Fig. 12. ↑ and ↓ represent larger and smaller is better, respectively.

Single dunhuang mural image	Methods	TransCNN-HAE	BOWN	DT	SIME-BINet (Ours)
(a)	PSNR ↑	23.7533	23.0743	22.0725	**24.3906**
	SSIM ↑	0.8976	0.8941	0.8810	**0.9197**
	LPIPS ↓	0.0809	0.0883	0.1045	**0.0637**
(b)	PSNR ↑	24.5860	24.5386	22.8703	**24.7911**
	SSIM ↑	0.8938	0.9082	0.8746	**0.9145**
	LPIPS ↓	0.0621	0.0637	0.0877	**0.0557**
(c)	PSNR ↑	**21.8612**	21.7681	19.8116	21.6852
	SSIM ↑	0.8181	0.8333	0.7878	**0.8388**
	LPIPS ↓	0.1081	0.1174	0.1281	**0.1037**
Average	PSNR ↑	23.4002	23.127	21.5848	**23.6223**
	SSIM ↑	0.8698	0.8785	0.8478	**0.8910**
	LPIPS ↓	0.0837	0.0898	0.1068	**0.0744**

4.2.2. Quantitative comparisons

Peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) (Wang et al., 2004), and learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018) are used to quantitatively evaluate SIME-BINet. PSNR indicates the quality of inpainted image. SSIM represents the structural similarity between repaired image and original image, and LPIPS learns image perceptual features using a deep learning model to quantify the similarity of images. In addition, higher values of PSNR and SSIM metrics are better, and lower values of LPIPS metrics are better.

The quantitative comparison results of SIME-BINet, TransCNN-HAE, BOWN, and DT are shown in Table 2. SIME-BINet achieves the best results on both pixel-level evaluation metrics PSNR and SSIM, and advanced visual evaluation metric LPIPS, by analyzing Table 2. Specifically, SIME-BINet shows remarkable improvements with its LPIPS being 32.98 % higher than DT, while improving PSNR and SSIM by 8.78 % and

Table 4
Quantitative results of ablation experiments. (a) refers to model with standard transformer block. (b) refers to IETSBlock with prediction mask. (c) Denotes IETSBlock which introduces prediction mask and self-information. Higher PSNR and SSIM values are better, lower LPIPS values are better.

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
(a) Base-TSBlock	24.7261	0.8759	0.0779
(b) Add IETSBlock with mask	24.8935	0.879	0.0764
(c) Add IETSBlock with enhanced information	**25.0997**	**0.8837**	**0.0701**

4.83 %, respectively. Notably, comparative experiments show that DT underperforms both TransCNN-HAE and BOWN in mural image processing. This performance gap can be primarily attributed to the inherently higher complexity of mural textures compared to natural images, which makes the multi-stage architecture of DT particularly susceptible to inter-stage interference during feature propagation. Furthermore, quantitative analysis is also conducted on sample images from Fig. 12, with the results presented in Table 3. The proposed SIME-BINet exhibits significant superiority, outperforming other methods by achieving the best results in both human vision (LPIPS) and structural (SSIM). These quantitative results are well corroborated by the visual quality observed in corresponding output images.

4.3. Ablation studies

SIME-BINet is decomposed through ablation experiments to verify the effectiveness of module design. Firstly, IESTBlock is replaced by TSBlock in SIME-BINet and used as a benchmark. Secondly, IETSBlock with only prediction mask enhancement is introduced. Finally, IETS-Block with prediction mask and self-information is introduced. Comparative experimental results for PSNR, SSIM, and LPIPS are shown in Table 4. As shown in Fig. 13, localized patch artifacts are present in repaired results when only TSBlock is employed. Then, patch artifacts are mitigated after the introduction of prediction mask enhancement. Finally, local information is further refined after the introduction of self-information enhancement, which is confirmed by the trend of visual evaluation metric LPIPS in Table 4. The above analysis validates the effectiveness of enhanced information and IETSBlock.

To explore the effects of self-information and prediction mask in the proposed blind inpainting method, the features of same decoding layer are analyzed when different enhanced information is used. In the case of without enhanced information, damaged and undamaged areas are not distinguished in the inpainting process. This leads to the interference from damaged areas with undamaged areas when repairing complex images, as shown in the second row of Fig. 14 (a). When only the prediction mask is introduced as enhanced information, the interference from damaged areas is reduced. However, the excessive focus on the regions to be repaired leads to a loss of detail in undamaged areas, as shown in Fig. 14 (b). After introducing self-information enhancement, the features of undamaged areas are more clearly preserved while repairing the damaged regions, as shown in Fig. 14 (b) and 14 (c).

4.4. Computational complexity analysis

Computational complexity is provided to analyze the usage restrictions of SIME-BINet. As shown in Table 5, the details include

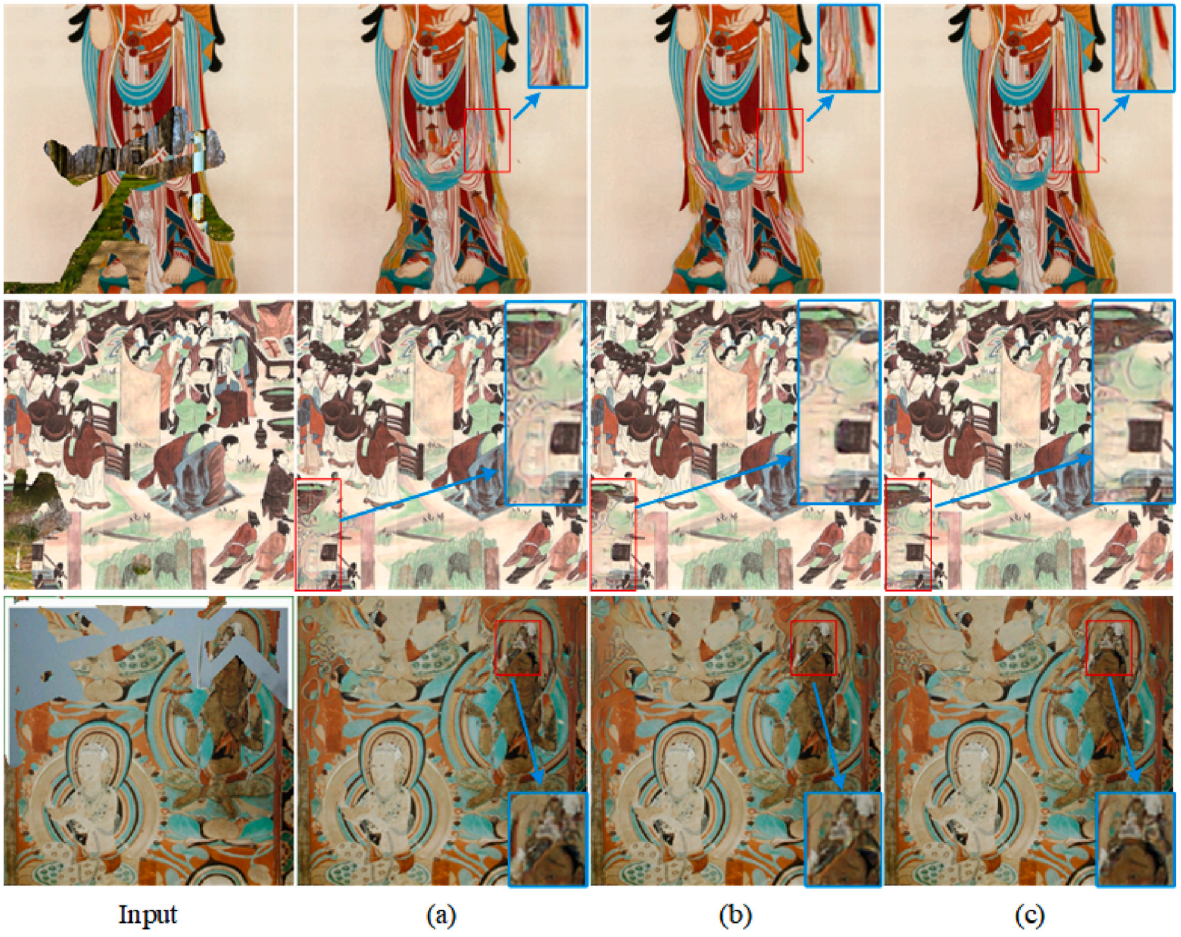


Fig. 13. Qualitative results of ablation experiments. (a) refers to Base-TSBlock. (b) refers to add IETSBlock with mask. (c) refers to add IETSBlock with self-prior. Please download the supplementary materials for more details.

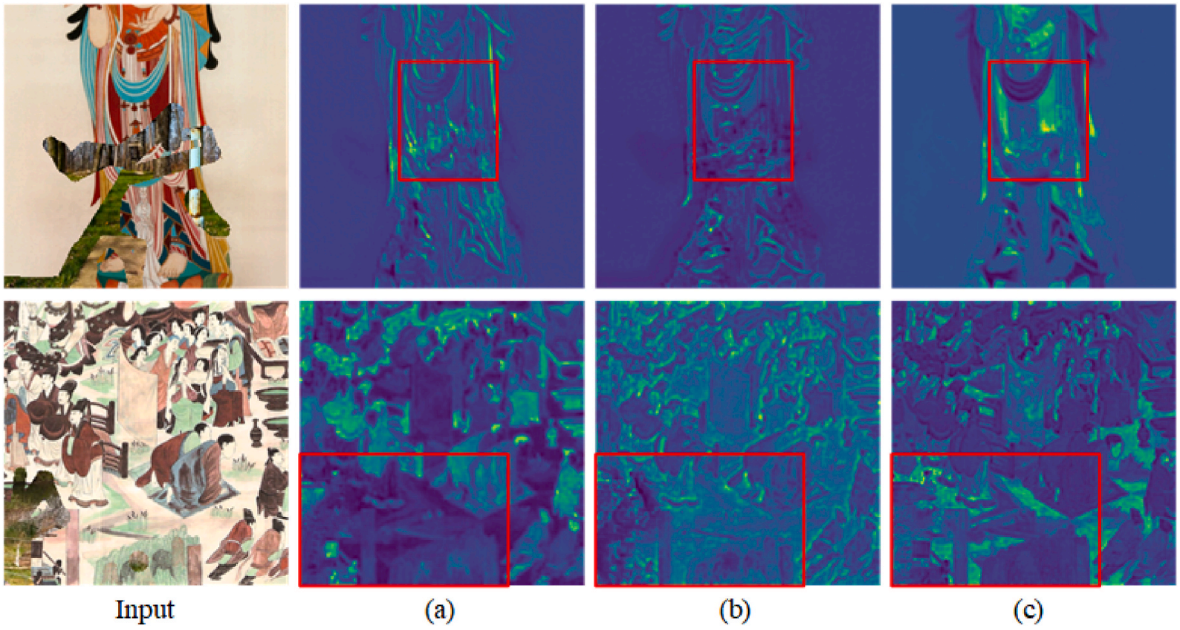


Fig. 14. Feature analysis of enhanced information. (a) refers to without enhanced information. (b) refers to adding a prediction mask as enhanced information. (c) refers to the addition of self-information as enhanced information. Please download the supplementary materials for more details.

Table 5
Computational complexity of SIME-BINet.

SIME-BINet	
Parameters	18.53 M
FLOPs	50.95 G
Infer. Time/per image	0.06 s
Training time	12 h 30 m
Model Size	70.72 MB
Maximum allocated memory	8.38 GB

parameters, flops, test time of single image (Infer. Time/per image), model size, and maximum allocated memory. By analyzing the data in Table 5, the average processing time for a single image is 0.06 s, which equates to 16.67 frames per second (fps). This is still a significant gap compared to the 24 fps of standard movies, indicating substantial potential for improvement in real-time inpainting. The storage space required for model saving is relatively small, at about 70.72 MB. The maximum peak memory usage during model running is 8.38 GB, indicating that SIME-BINet can be deployed on consumer-level GPUs with memory greater than 8.38 GB, such as NVIDIA GeForce GTX 1080Ti and NVIDIA GeForce RTX 3060. Please note that the data in Table 5 were measured using 256 × 256 input images, and the inference time is the average of 50 images tested with an NVIDIA GeForce RTX 4090.

4.5. Test of real damage masks

Current mural image inpainting methods are mainly trained and tested by simulated damage. This section presents real damage mask inpainting experiments to evaluate SIME-BINet more objectively, as shown in Fig. 15. In general, blind inpainting methods can repair real damage mask mural images without damaged mask annotation, but inpainting results are insufficient. Specifically, the repair results of SIME-BINet have no obvious artifacts and have better visual effects, the repair results of TransCNN-HAE and BOWN methods have local plaque artifacts, and the repair results of DT have structural disorder, as shown in Fig. 15. And above results are consistent with problems in simulation masks repair. At the same time, real damage mask experiments also show limitations of SIME-BINet, although repair results have better visual effects compared to other methods, but still shortcomings in the recovery of local complex details.

4.6. Discussion of limitations

The proposed method has limitations in repairing complex textures and structural damage. As shown in Fig. 16, the internal details of damaged areas remain insufficiently clear, although the proposed method achieves competitive performance in both comprehensive metrics and visual effects. This limitation can be attributed to the complexity of local structural textures and insufficient datasets. Therefore, further research on repair objects with complex textures and expansion dunhuang mural image datasets are essential to improve SIME-BINet performance. At the same time, reducing model resource consumption and achieving real-time repair are also important research directions for the future. Additionally, leveraging the powerful generation capabilities of current large-scale models to enhance generalization performance is another key area for future research.

5. Conclusion

This paper proposed a novel end-to-end self-information and prediction mask enhanced blind inpainting network (SIME-BINet) for dunhuang mural images. SIME-BINet introduces enhanced features composed of valid information from undamaged areas and a prediction mask through information enhancement transformer module, which significantly improves the visual results of blind inpainting method. Simulated and real mask experiments show the structural and visual superiority of SIME-BINet.

CRediT authorship contribution statement

Jiahao Meng: Writing – original draft, Software, Methodology, Conceptualization. **Weirong Liu:** Writing – review & editing, Resources, Methodology, Conceptualization. **Changhong Shi:** Validation, Investigation, Data curation. **Zhijun Li:** Validation, Software, Data curation. **Jie Liu:** Writing – review & editing, Supervision.

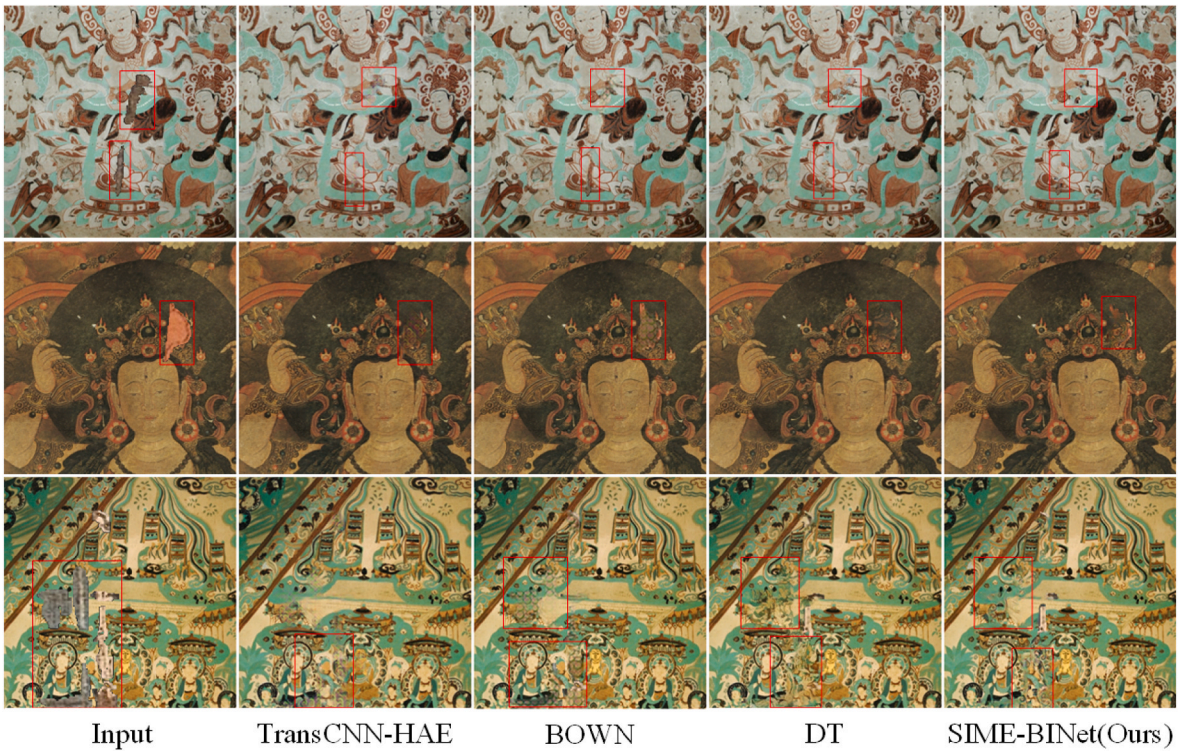


Fig. 15. Comparison test of real damaged mask. Please download the supplementary materials for more details.

Fig. 16. Limitations of SIME-BINet.

Declaration of competing 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.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62261032, the Central Guidance for Local Development Projects, and the Key Talent Project of Gansu Province.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.engappai.2025.111769.

Data availability

The data, code, and pre-training model of SIME-BINet will be released on Github (https://github.com/IPCSRG/SIME-BINet) to further validate the authenticity of the experiments after acceptance.

References

Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B., 2009. PatchMatch: a randomized correspondence algorithm for structural image editing. ACM Trans. Graph. 28, 24.
Bertalmio, M., Vese, L., Sapiro, G., Osher, S., 2003. Simultaneous structure and texture image inpainting. IEEE Trans. Image Process. 12, 882–889.
Chan, T.F., Shen, J., 2001. Nontexture inpainting by curvature-driven diffusions. J. Vis. Commun. Image Represent. 12, 436–449.
Chen, G., Zhang, G., Yang, Z., Liu, W., 2023. Multi-scale patch-GAN with edge detection for image inpainting. Appl. Intell. 53, 3917–3932.
Chen, L., Lu, X., Zhang, J., Chu, X., Chen, C., 2021. Hinet: half instance normalization network for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 182–192.
Chen, Y., Xia, R., Yang, K., Zou, K., 2024a. DNNAM: image inpainting algorithm via deep neural networks and attention mechanism. Appl. Soft Comput. 154, 111392.
Chen, Y., Xia, R., Yang, K., Zou, K., 2024b. Image inpainting algorithm based on inference attention module and two-stage network. Eng. Appl. Artif. Intell. 137, 109181.
Chen, Y., Xia, R., Yang, K., Zou, K., 2024c. MICU: image super-resolution via multi-level information compensation and U-net. Expert Syst. Appl. 245, 123111.
Criminisi, A., Pérez, P., Toyama, K., 2004. Region filling and object removal by exemplar-based image inpainting. IEEE Trans. Image Process. 13, 1200–1212.
Deng, X., Yu, Y., 2023. Ancient mural inpainting via structure information guided two-branch model. Heritage Science 11, 131.
Gao, L., Zhan, H., Sheng, V.S., 2024a. Exploring task-specific dimensions in word embeddings through automatic rule learning. In: Proceedings of the International Conference on Artificial Neural Networks. Springer, pp. 199–214.
Gao, T., Wen, Y., Zhang, J., Chen, T., 2024b. A novel dual-stage progressive enhancement network for single image deraining. Eng. Appl. Artif. Intell. 128, 107411.
Gatys, L.A., Ecker, A.S., Bethge, M., 2016. Image style transfer using convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, Las Vegas, USA, pp. 2414–2423.
Hao, Q., Qiaomei, Z., 2010. Master Ji Xianlin's academic contribution to dunhuang study. Hundred Schools in Arts 26, 219–220+231.
Huang, L., Huang, Y., 2023. DRGAN: a dual resolution guided low-resolution image inpainting. Knowl. Base Syst. 264, 110346.
Iizuka, S., Simo-Serra, E., Ishikawa, H., 2017. Globally and locally consistent image completion. ACM Trans. Graph. 36, 1–14.
Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Springer, Amsterdam, The Netherlands, pp. 694–711.
Kumar, P., Gupta, V., Grover, M., 2024. Dual attention and channel transformer based generative adversarial network for restoration of the damaged artwork. Eng. Appl. Artif. Intell. 128, 107457.
Le Meur, O., Gautier, J., Guillemot, C., 2011. Examplar-based inpainting based on local geometry. In: Proceedings of the IEEE International Conference on Image Processing. IEEE, pp. 3401–3404.
Li, C.-Y., Lin, Y.-Y., Chiu, W.-C., 2023. Decontamination transformer for blind image inpainting. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing IEEE, pp. 1–5.
Li, L., Zou, Q., Zhang, F., Yu, H., Chen, L., Song, C., Huang, X., Wang, X., Line drawing guided progressive inpainting of mural damages. arXiv preprint arXiv:2211.06649.
Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., Li, J., 2020. Dice loss for data-imbalanced nlp tasks. In: Proceedings of the the Association for Computational Linguistics, pp. 465–476.
Li, X., Wang, Z., Chen, C., Tao, C., Qiu, Y., Liu, J., Sun, B., 2024. Semid: blind image inpainting with semantic inconsistency detection. Tsinghua Sci. Technol. 29, 1053–1068.
Liu, G., Reda, F.A., Shih, K.J., Wang, T.-C., Tao, A., Catanzaro, B., 2018. Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European Conference on Computer Vision, pp. 85–100.
Liu, W., Shi, Y., Li, J., Wang, J., Du, S., 2023. Multi-stage progressive reasoning for dunhuang murals inpainting. In: Proceedings of the IEEE International Conference on Pattern Recognition and Machine Learning IEEE, pp. 211–217.
Meng, J., Liu, W., Shi, C., Li, Z., Liu, C., 2024. Degression receptive field network for image inpainting. Eng. Appl. Artif. Intell. 138, 109397.
Mosleh, A., Bouguila, N., Hamza, A.B., 2013. Automatic inpainting scheme for video text detection and removal. IEEE Trans. Image Process. 22, 4460–4472.
Mosleh, A., Sola, Y.E., Zargari, F., Onzon, E., Langlois, J.P., 2017. Explicit ringing removal in image deblurring. IEEE Trans. Image Process. 27, 580–593.
Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A., 2016. Context encoders: feature learning by inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, Las Vegas, USA, pp. 2536–2544.
Peng, C., Chellappa, R., 2023. PDRF: progressively deblurring radiance field for fast scene reconstruction from blurry images. Proceedings of the AAAI Conference on Artificial Intelligence, pp. 2029–2037.
Phutke, S.S., Kulkarni, A., Vipparthi, S.K., Murala, S., 2023. Blind image inpainting via omni-dimensional gated attention and wavelet queries. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1251–1260.
Ren, H., Sun, K., Zhao, F., Zhu, X., 2024. Dunhuang murals image restoration method based on generative adversarial network. Heritage Science 12, 39.
Ruby, U., Yendapalli, V., 2020. Binary cross entropy with deep learning technique for image classification. International Journal of Advanced Trends in Computer Science 9.
Schmidt, A., Madhu, P., Maier, A., Christlein, V., Kosti, R., 2022. ARIN: adaptive resampling and instance normalization for robust blind inpainting of dunhuang cave paintings. In: Proceedings of the International Conference on Image Processing Theory, Tools and Applications IEEE, pp. 1–6.
Shen, J., Chan, T.F., 2002. Mathematical models for local nontexture inpaintings. SIAM J. Appl. Math. 62, 1019–1043.
Simonyan, K., Zisserman, A., Very Deep Convolutional Networks for large-scale Image Recognition. arXiv preprint arXiv:1409.1556.
Sun, W., Chen, Z., 2020. Learned image downscaling for upscaling using content adaptive resampler. IEEE Trans. Image Process. 29, 4027–4040.
Wang, H., Jiang, L., Liang, R., Li, X.-X., 2017. Exemplar-based image inpainting using structure consistent patch matching. Neurocomputing 269, 90–96.
Wang, J., Yuan, C., Li, B., Deng, Y., Hu, W., Maybank, S., 2023. Self-prior guided pixel adversarial networks for blind image inpainting. IEEE Trans. Pattern Anal. Mach. Intell. 45, 12377–12393.
Wang, Y., Chen, Y.-C., Tao, X., Jia, J., 2020. Vcnet: a robust approach to blind image inpainting. In: Proceedings of the European Conference on Computer Vision. Springer, pp. 752–768.
Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13, 600–612.
Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S., 2018. Generative image inpainting with contextual attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5505–5514.
Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 586–595.
Zhang, X., Zhai, D., Li, T., Zhou, Y., Lin, Y., 2023. Image inpainting based on deep learning: a review. Inf. Fusion 90, 74–94.
Zhao, F., Ren, H., Sun, K., Zhu, X., 2024a. GAN-Based heterogeneous network for ancient mural restoration. Heritage Science 12, 418.
Zhao, H., Gu, Z., Zheng, B., Zheng, H., 2022. Transcnn-hae: transformer-Cnn hybrid autoencoder for blind image inpainting. In: Proceedings of the ACM International Conference on Multimedia, pp. 6813–6821.
Zhao, H., Wang, Y., Gu, Z., Zheng, B., Zheng, H., 2024b. Context-aware mutual learning for blind image inpainting and beyond. Expert Syst. Appl., 126224
Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A., 2017. Places: a 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. 40, 1452–1464.
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, 1–25.

Jiahao Meng is a Ph.D. candidate at the College of Electrical and Information Engineering, Lanzhou University of Technology. His main research interests are image processing and pattern recognition.

Weirong Liu received the Ph.D. degree, from Hunan University, Changsha, China, in 2014. 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.

Changhong Shi is a Ph.D. candidate at the College of Electrical and Information Engineering, Lanzhou University of Technology. Her main research interests are image processing and pattern recognition.

Zhijun Li is a Ph.D. candidate at the College of Electrical and Information Engineering, Lanzhou University of Technology. His main research interests are image processing and pattern recognition.

Jie Liu is currently a Senior Engineer at the College of Electrical and Information Engineering, Lanzhou University of Technology. Her principal research interests are on topics related to image processing and pattern recognition.

Journal of Southeast University(English Edition)
东南大学学报(英文版)
ISSN 1003-7985,CN 32-1325/N

《Journal of Southeast University(English Edition)》网络首发论文

题目：基于多感受野与破损模式动态匹配的图像修复方法(英文)
作者：孟家豪，刘微容，史长宏，李治俊，刘婕
收稿日期：2025-07-03
网络首发日期：2025-10-14
引用格式：孟家豪，刘微容，史长宏，李治俊，刘婕．基于多感受野与破损模式动态匹配的图像修复方法(英文)[J/OL]. Journal of Southeast University(English Edition). https://link.cnki.net/urlid/32.1325.N.20251014.1441.002

Published online：2025-10-14 15:30:18 URL：https://link.cnki.net/urlid/32.1325.N.20251014.1441.002

Multi-receptive fields and dynamic matching of damaged patterns for image inpainting

MENG Jiahao, LIU Weirong, SHI Changhong, LI Zhijun, LIU Jie

（School of Automation and Electrical Engineering, Lanzhou University of Technology, Lanzhou 730050, China）

Abstract: Existing image inpainting models are primarily designed based on large receptive field, and utilization of different receptive fields is realized by refinement networks. However, the adaptation between different receptive fields and image damage patterns is ignored, which causes problems such as artifacts and semantic information confusion in repaired images. To address the problems of artifacts and semantic information confusion, inspired by different sensitivities of different receptive fields to inpainting different damaged patterns, this paper proposes an image inpainting method based on multi-receptive fields and dynamic matching of damaged patterns. Firstly, multi-receptive field feature groups are extracted by parallel filter banks. Secondly, dynamic matching relationship between receptive field and damaged pattern is constructed by dynamic weighting and screening of multi-receptive field feature groups under constraint of mask image. In addition, a decoder based on fast fourier convolution is designed to enhance fusion ability of global context features when recovering high-dimensional features to low-dimensional images during decoding process. Comparative experimental results show that proposed method can achieve better subjective and objective inpainting results on three public datasets: Paris StreetView, CelebA-HQ, and Places2.

Key words: image inpainting; generative adversarial networks; multi-receptive fields; dynamic matching of damaged patterns; decoder with fast fourier convolutional

In recent years, with the development of deep learning, the performance and applicability of computer vision methods have been significantly improved. Deep Neural Networks are widely applied in areas such as object tracking and recognition[1], parameter detection[2], and image inpainting[3].

Image inpainting is one of basic tasks in computer vision, and inpainting infer and supplement content of missing region by using information of non-missing region in damaged image. In recent years, image inpainting based on deep learning has been widely used in image editing and cultural relic image restoration. Before popularity of deep learning, image inpainting was mainly dominated by partial differential equation basedand block matching based methods[4]. Image inpainting based on partial differential equation propagates gradient information of undamaged region to inpainted region by diffusion along direction orthogonal to gradient. Block matching based approach searches for blocks in undamaged region that are similar to damaged region and then copies them to damaged region for filling. Partial differential equation and block matching methods have achieved good

Received：2025-07-03

Biography：Meng Jiahao（1993—），male，PhD candidate；Liu Weirong（Corresponding author），male，doctor，professor，Ph. D. Supervisor，liuwr@lut.edu.cn.

Foundation item：The National Nature Science Foundation of China under Grant（62261032）；The Central Government Guiding Funds for Local Science and Technology Development Program（25ZYJA026）.

results for inpainting with high similarity between damaged and undamaged area (such as white wall with high consistency). However, due to the lack of flexibility, they cannot achieve good results in case of large continuous damaged or undamaged area with less reference information.

In recent years, researchers have gradually introduced convolutional neural networks and generative adversarial networks to image inpainting because of wide adaptability of deep networks in computer vision[5, 6]. In general, these methods can be categorized into single receptive field[7, 8] and multi-receptive field networks[9-14] according to network structure. Single receptive field networks typically use single codec as generator, and usually have problems such as local texture blur and artifacts because of large and single receptive fields. To increase receptive field, some researchers have proposed multi-receptive field networks from coarse-to-fine, and such models are dominated by two-stage models consisting of large and small receptive field networks. Multi-receptive field network has achieved better repair results compared with single receptive field network, but due to adaptation between multi-receptive fields and image damage patterns is not considered in network design, above problems still exist.

Recently, Zheng et al.[12] proposed that convolutions with large receptive fields cannot get rid of the interaction of neighboring pixels, and small receptive fields can limit influence of mask region to neighboring pixels. Inspired by idea of Zheng, some experiments were conducted to analyze the effect of receptive field size on image inpainting based on pure convolutional codec. It can be concluded that different receptive fields have different repair sensitivities to different damage patterns. However, existing image inpainting models have problems of artifacts and confusing semantic information because they are mostly designed based on large receptive fields and ignore adaptability between different receptive fields and image damage patterns.

To address above problems, an image inpainting method based on dynamic matching of multi-receptive fields and damaged patterns is proposed, with extraction and utilization of multi-receptive field features as main design idea based on previous analysis. The proposed method mainly consists of three parts, Multi-RF-Feature Generation Module (MRFG), Feature Mapping and Selection Module (FMS) and decoder with fast fourier convolutional. Specifically, MRFG module with filter bank is first designed as encoder to extract unified scale multi-receptive field feature group of damaged image. Secondly, in order to obtain hybrid receptive field feature group that adapted to current damage pattern, FMS module maps damaged mask image into learnable one-dimensional score features at first, and then weights obtained score features with unified scale multi-receptive field features. In addition, a decoder based on fast fourier convolution is further constructed to utilize global contextual feature fusion capability of Fast Fourier Convolution (FFC) to solve problem that pure convolutional neural networks have certain limitations in global contextual feature fusion[12]. The main contributions are as follows:

1. A novel single-stage end-to-end network for image inpainting is proposed. The network innovatively takes dynamic matching relationship between multi-receptive fields and image damage

patterns as breakthrough point, and uses learnable latent feature mapping of masks and the extraction and rescreening mechanism of multi-receptive field feature groups to establish matching relationship.

2. A multi-receptive field feature generation module is proposed. This module uses designed filter bank as encoder to realize extraction of multi-receptive field features of unified scale, which is different from existing multi-receptive field models that utilize different receptive fields through a series of multi-stage networks.

3. A new damaged feature mapping and dynamic matching module is proposed. This module takes mask image as constraint condition and constructs dynamic matching relationship between receptive field and damaged pattern through weighted sum screening of unified scale multi-receptive field feature group, which is different from existing methods that ignore adaptation between receptive field and damaged pattern.

1 Methods

1.1 Overview

As an important attribute in deep neural networks, Receptive Field (RF) has been widely concerned in downstream tasks of computer vision such as object detection, but it still has not received enough attention in image inpainting[12]. As shown in Fig. 1, small receptive field is susceptible to interference of neighboring pixels when repair damaged image with relatively large area, while large receptive field tends to ignore detail information of damaged area when repair relatively small area of damage.

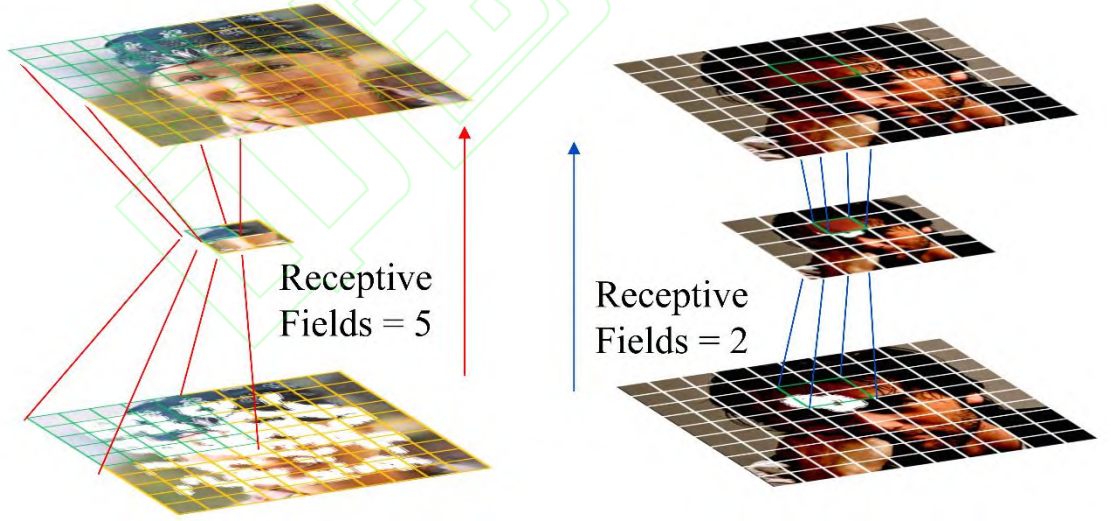


Fig. 1 Schematic representation the role of receptive fields in image inpainting tasks

Inspired by above analysis, a novel inpainting network architecture with dynamic matching of multi-receptive fields and damaged patterns is proposed to address artifacts and semantic information confusion because of current inpainting network lacks the matching relationship between receptive fields and damaged patterns. Specifically, the proposed model is single-stage network composed of single codec. As shown in Fig. 2, the 0 in figure represents damaged area, and

nonzeros numbers represent undamaged areas. By observing the green region reveals that features with receptive field size 2 fail to capture any effective information from their coverage area. In contrast, within the red bounding box, features with receptive field size 3 can generate effective high-dimensional features by successfully integrating information from both the damaged region and its surrounding context. Since different receptive fields exhibit varying sensitivities to different damages, the proposed model constructs an encoder consisting of Multi-RF-Feature Generation Module (MRFG) and Feature Mapping and Selection Module (FMS) to establish the matching relationship between damage and receptive field, as shown in blue and purple areas in Fig. 3. Meanwhile, to improve ability of global and local feature fusion in decoding process, Fast Fourier Convolution (FFC) module is introduced into decoder, as shown in green area in Fig. 3. The specific process of the proposed model is as follows.

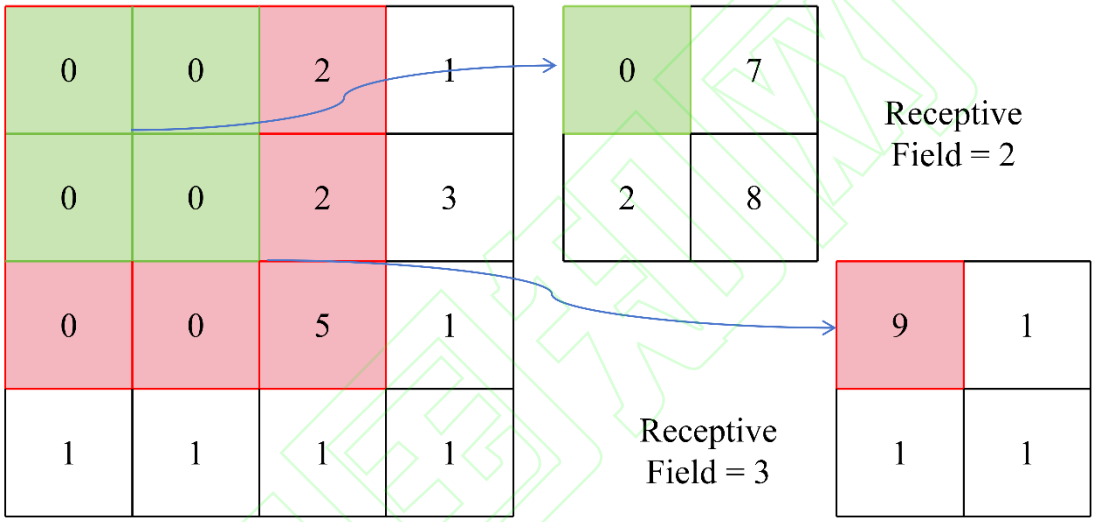


Fig. 2 Schematic diagram of feature computation with different receptive fields

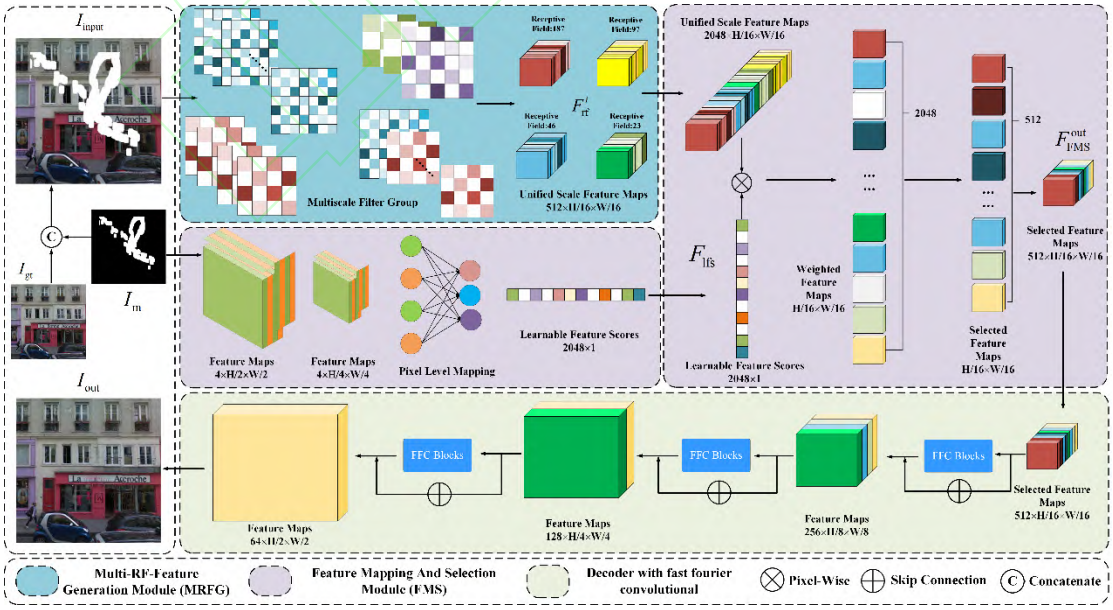


Fig. 3 Architecture of proposed multi-receptive fields and damage pattern dynamic matching network

As shown in Fig. 3, given original image I_{gt} and mask I_m, input I_{input} of multi-receptive field feature generation module is as follows:

$$I_{dmg} = I_{gt} \odot (1 - I_m) \tag{1}$$

$$I_{input} = Cat(I_{dmg}, I_m) \tag{2}$$

where I_{dmg} denotes damaged image and $Cat(\cdot)$ represents concatenation of damaged image I_{dmg} and mask image I_m by channel dimension.

Secondly, MRFG module extracts feature group F_{rf}^i with uniform scale and different receptive fields. Meanwhile, mask image I_m is mapped into learnable one-dimensional weighted features F_{lfs} by feature mapping module.

$$F_{rf}^i = Enc_{MRFG}^i(I_{input}) \quad i = 1, 2, \cdots, n \tag{3}$$

$$F_{lfs} = Enc_{FMS}^M(I_m) \tag{4}$$

where $Enc_{MRFG}^i(\cdot)$ represents i th subnetwork with different receptive fields, F_{rf}^i F_{rf}^i represents output features of i th subnetwork, and $Enc_{FMS}^M(\cdot)$ denotes feature mapping part in FMS module.

Then, the feature group F_{rf}^i and weighting feature F_{lfs} are used as inputs to be fused and computed by damage feature matching module to obtain feature F_{FMS}^{out} that has a higher match with current image damage pattern:

$$F_{FMS}^{out} = Enc_{FMS}^S(Cat(F_{rf}^i), F_{lfs}) \quad i = 1, 2, \cdots, n \tag{5}$$

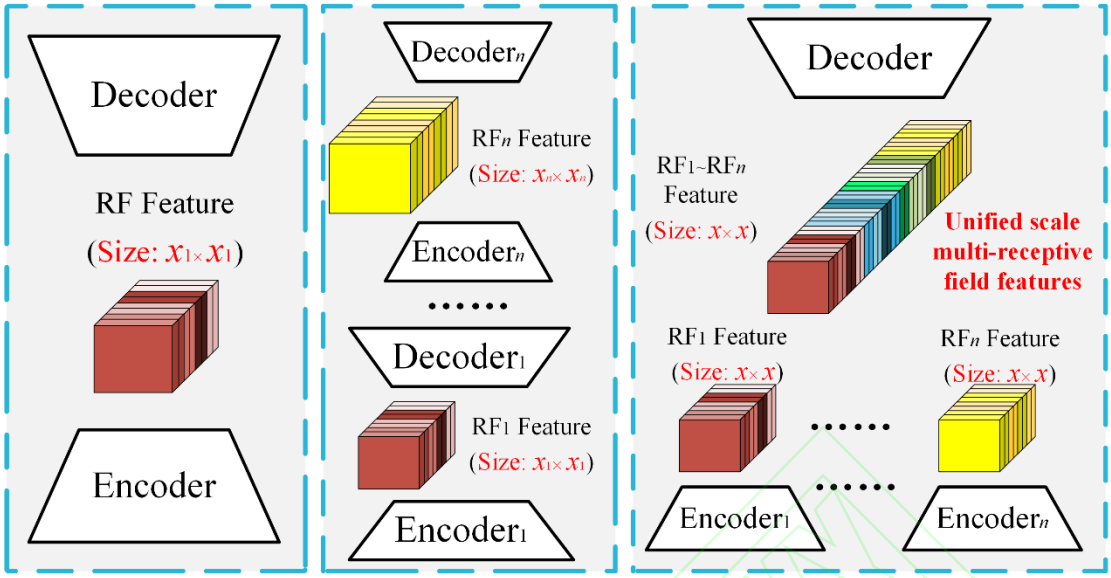
where $Enc_{FMS}^S(\cdot)$ represents feature matching part in FMS module.

Finally, F_{FMS}^{out} is used as input feature and decoded by decoder $Dec_{FFC}(\cdot)$, which was designed based on fast fourier convolution.

$$I_{out} = Denc_{FFC}(F_{FMS}^{out}) \tag{6}$$

where I_{out} is final repaired image.

1.2 Multi-RF-Feature generation module (MRFG)



(a) SRFN (b) MRFN (c) MRFG

Fig. 4 Comparison diagram between the MRFG module and current network structure. SRFN represents single receptive field network and MRFN represents multi-receptive field network

Table 1 Structure of MRFG module. RF represents receptive field. k, s, p denote kernel size, stride, and padding, respectively

Subnetwork	Filter(k, s, p)	RF	Output
Enc^1_{MRFG}	(3，2，1), (3，2，1), (5，4，1)	23×23	16×16
Enc^2_{MRFG}	(4，2，1)×4	46×46	16×16
Enc^3_{MRFG}	(7，2，5), (4，2，5), (4，2，1)×3	97×97	16×16
Enc^4_{MRFG}	(7，2，1)×2, (7，2，0), (7，1，1)×3	187×187	16×16

Multi-receptive field feature generation module (MRFG) is designed to enhance multi-receptive field feature extraction capability of network and obtain high-dimensional feature groups with different sizes of receptive fields. Current image inpainting methods usually use codecs with different receptive fields to realize application of multi-receptive field features, as shown in Fig. 4. Single receptive field networks are usually single-stage network, as shown in Fig. 4 (a). Multi-receptive field networks realizes utilization of multiple receptive fields by concatenating multiple codecs, as shown in Fig. 4 (b). Unlike existing methods, the MRFG module constructs feature extraction network composed of multiple sub encoders through filter bank to realize extraction of

multi-scale receptive field features, and then provides feature set to be selected for dynamic matching of different receptive field features with image damage patterns. Specific structure of MRFG is shown in Table 1. Specifically, with I_{input} as input, separate filter banks with different receptive field feature extraction capabilities are used for feature extraction:

$$F_{rf}^{i} = Enc_{\text{MRFG}}^{i}(I_{\text{input}}) \quad i \in \{1,2,3,4\} \tag{7}$$

where F_{rf}^{i} is group of unified scale features with different receptive fields.

1.3 Feature mapping and selection module (FMS)

Inspired by different sensitivity of different receptive field features to damage patterns, damage feature mapping and selection module (FMS) is innovatively designed to overcome the mixing effect on simple superposition of multiple receptive field features, and establishes dynamic matching relationship between image damage pattern and multi-receptive field feature groups. FMS module mainly consists of two parts: damage feature mapping and feature matching. Specifically, damage feature mapping part, mainly through $PLM(\cdot)$ (Pixel Level Mapping, PLM) maps damage mask image I_{m} into learnable one-dimensional weight features:

$$F_{\text{lfs}} = Enc_{\text{FMS}}^{\text{M}}(I_{\text{m}}) = PLM(Conv(Conv(I_{\text{m}}))) \tag{8}$$

After obtaining F_{lfs}, F_{rf}^{i} and F_{lfs} are simultaneously input to damage feature matching part, and features in F_{rf}^{i} are weighted by weighting feature F_{lfs} to achieve effect of suppressing low-frequency invalid features and enhancing high-frequency valid features, and then the weighted multi-receptive field feature set $F_{\text{rf}}^{\text{weighted}}$ is obtained. Finally, $F_{\text{rf}}^{\text{weighted}}$ is filtered by operation $Selection(\cdot)$ to select high-frequency effective features and obtain mixed receptive field feature set that is applicable to current damage pattern after filtering:

$$F_{\text{rf}}^{\text{weighted}} = Cat(F_{\text{rf}}^{1}, F_{\text{rf}}^{2}, F_{\text{rf}}^{3}, F_{\text{rf}}^{4}) \otimes F_{\text{lfs}} \tag{9}$$

$$\begin{aligned} F_{\text{FMS}}^{\text{out}} &= Enc_{\text{FMS}}^{\text{S}}(Cat(F_{\text{rf}}^{i}), F_{\text{lfs}}) \\ &= Selection(Leaky\,\text{Re}\,LU(Norm(F_{\text{rf}}^{\text{weighted}})) \\ &\qquad i = 1,2,\cdots,n \end{aligned} \tag{10}$$

where $\otimes$ represents pixel-wise dot multiplication.

1.4 Decoder with fast fourier convolution

Thanks to the ability of fast fourier convolution (FFC) module to perform fusion of global contextual features without changing feature sizes by means of spectral transform, fast fourier convolution-based decoder is designed to overcome the limitations of pure convolutional decoders in global context feature fusion.

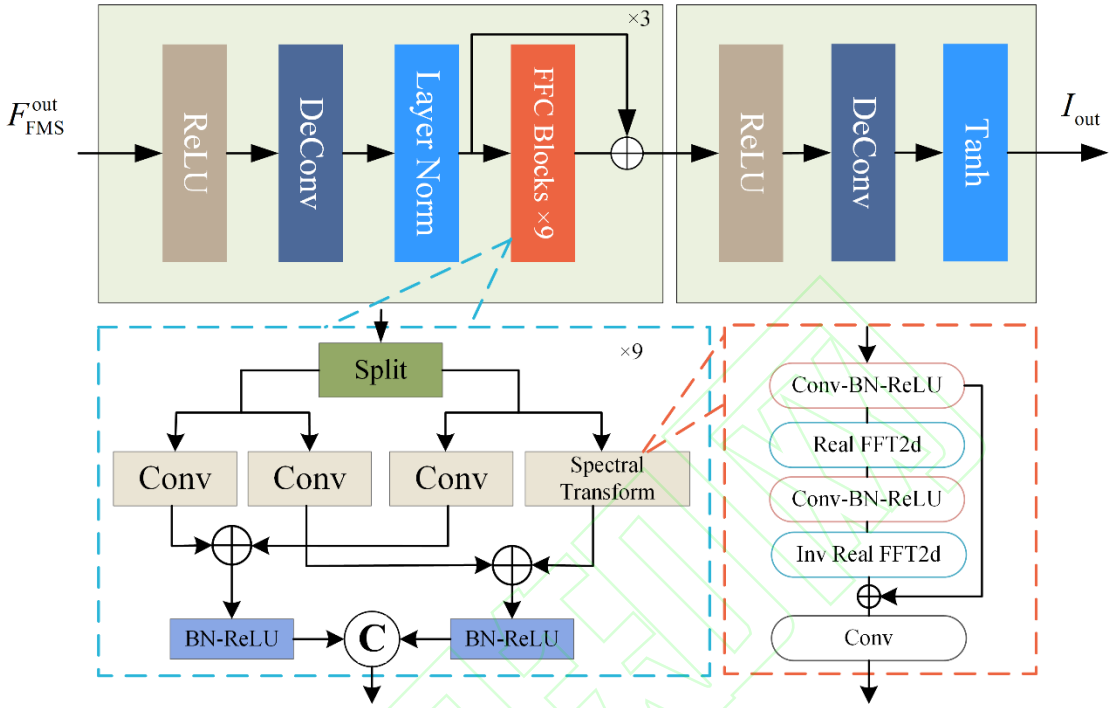


Fig. 5 Decoder network architecture based on Fast Fourier convolution

Decoder network architecture is shown in Fig. 5 and consists of three upsample groups with FFC modules and output head. Where encoder output F_{FMS}^{out} is used as input feature of decoder, output $F_{decoder}^{out}$ of single upsample group can be represented as:

$$F_{decoder} = Norm(DeConv(ReLU(F_{FMS}^{out}))) \tag{11}$$

$$F_{decoder}^{out} = F_{decoder} \oplus FFC^{j}(F_{decoder}) \qquad j = 1, 2, \cdots, m \tag{12}$$

where $ReLU()$ is activation function, $DeConv()$ denotes deconvolution, and $Norm()$ is normalization.

Then, $F_{decoder}^{out}$ is used as the input of next upsample group and result $F_{decoder}^{out}$ is reorganized through output head to obtain final repaired image I_{out}.

$$I_{out} = Tanh(DeConv(ReLU(F_{decoder}^{out}))) \tag{13}$$

1.5 Loss function

Composite loss function is used to better train model. Firstly, weighted L1 loss is used for

pixel-level reconstruction to optimize training process and reduce training time as shown in following equation:

$$\ell_{\text{hole}} = \frac{1}{sum(I_{\text{m}})} \left\| (I_{\text{out}} - I_{\text{gt}}) \odot I_{\text{m}} \right\| \tag{14}$$

$$\ell_{\text{valid}} = \frac{1}{sum(1-I_{\text{m}})} \left\| (I_{\text{out}} - I_{\text{gt}}) \odot (1-I_{\text{m}}) \right\| \tag{15}$$

$$\ell_{\text{L1}} = \ell_{\text{valid}} + \lambda_{\text{h}} \cdot \ell_{\text{hole}} \tag{16}$$

Secondly, the total variation loss[15] was used as smoothing penalty term to reduce noise in generation process and make image smoother. Meanwhile, in order to better recover structure and texture details, perceptual loss[16] and style loss[17] are also introduced into loss function. The total loss function of generator is as follows.

$$\ell_G = \ell_{\text{L1}} + \lambda_{\text{tv}} \cdot \ell_{\text{tv}} + \lambda_{\text{perceptual}} \cdot \ell_{\text{perceptual}} + \lambda_{\text{style}} \cdot \ell_{\text{style}} \tag{17}$$

where λ_{h}, λ_{tv}, $\lambda_{\text{perceptual}}$, and λ_{style} are regularization parameters for each loss.

Finally, adversarial loss is introduced in training process to make distribution of repaired image I_{out} closer to real image I_{gt} in training process, and loss function is as follows:

$$\begin{aligned} \ell_D = \min_G \max_D \mathbb{E}_{I_{gt} \sim \mathbb{P}_r} [logD(I_{\text{gt}})] \\ - \mathbb{E}_{I_{out} \sim \mathbb{P}_g} [log(D(I_{\text{out}}))] \end{aligned} \tag{18}$$

where D represents discriminator, and structure is shown in Table 2.

Table 2 Discriminator structure table. IC represents input channel. OC represents output channel. k, s, p denote kernel size, stride, and padding, respectively

Layer	Activation	IC	OC	(k, s, p)
Conv1	LeakyReLU	3	64	(5，2，1)
Conv2	LeakyReLU	64	128	(5，2，1)
Conv3	LeakyReLU	128	256	(5，2，1)
Conv4	LeakyReLU	256	512	(5，1，1)
Conv5	-	512	1	(5，1，1)

In summary, proposed model is trained in end-to-end manner and final loss function is as follows:

$$\ell = \ell_G + \ell_D \tag{19}$$

2 Experiments

2.1 Experiment setup

2.1.1 Dataset and evaluation metrics

Three of most commonly used public datasets in image inpainting are used to evaluate the performance of proposed model.

Paris StreetView[18]: contains 15,000 images of buildings, trees, streets, and sky in Paris, France.

CelebA-HQ[19]: Contains 30,000 high-resolution face images.

Places2[20]: A large-scale scene recognition dataset containing 8 millions of images of various scenes.

The details of dataset are shown in Table 3. Among them, the Places2 dataset follows the literature[21] to randomly select 2,000 images from each of the first 20 categories as training set, which contains 40,000 images, and at the same time, 100 images are randomly selected from these 20 categories as test set, which contains 2,000 images.

Table 3 Train and test dataset details

Dataset	Train	Test	Total
Paris StreetView	14900	100	15000
CelebA-HQ	28000	2000	30000
Places2	40000	2000	42000

Peak signal-to-noise ratio (PSNR), the structural similarity index (SSIM)[22], Fréchet inception distance (FID)[23] and learned perceptual image patch similarity (LPIPS)[24] are used to evaluate proposed method. In addition, PSNR and SSIM are pixel-level evaluation metrics, while FID and LPIPS are related to high-level visual perception.

2.1.2 Comparison methods

In order to validation advancement of proposed method, quantitative and qualitative comparisons were conducted with six state-of-the-art (SOTA) methods in experimental analysis, and comparison methods are shown in Table 4.

Table 4 Comparison methods

Method	Venue & Year	Intro
MADF[25]	IEEE TIP'21	A Coarse-to-fine Image inpainting network with mask-aware dynamic filtering module and point normalization.
TFill[12]	CVPR'22	First two-stage inpainting network that introduces transformer to image inpainting tasks.
W-Net[9]	IEEE TMM'23	W-shaped two-stage inpainting network fusing texture spatial attention and structural channel excitation.
AOT[26]	IEEE TVCG'23	Aggregated contextual-transformation enhanced GAN-based model for image inpainting.
FCF[27]	AAAI'23	Fourier coarse-to-fine generator for general-purpose image inpainting.
SCAT[28]	WACV'23	Adversarial training framework for image inpainting with segmentation confusion adversarial training and contrastive learning.

2.1.3 Implement details

Adam optimizer is used in model training, where $\beta_1 = 0.5, \beta_2 = 0.999$ A decay learning strategy is used in training process, specifically, Epoch = 200, the initial learning rate of first 100 Epochs is set to be 0.0002, and the learning rate of last 100 Epochs is linearly decayed to 0. Regularization parameters of each loss function are $\lambda_{\text{h}} = 6$, $\lambda_{\text{tv}} = 0.1$, $\lambda_{\text{perceptual}} = 0.05$, and $\lambda_{\text{style}} = 120$ in training phase. Meanwhile, all images and mask sizes are 256×256 in all experiment. All models are trained on single NVIDIA GeForce RTX 4090 GPU.

2.2 Experiment results and analysis

Qualitative and quantitative comparison experiments are conducted to verify the effectiveness and advancement of the proposed method. Furthermore, the limitations are analyzed through large damage experiments. Note that official code and pre-trained models are used in all experiments.

2.2.1 Qualitative comparisons

Fig. 6, 7 and 8 show qualitative comparison results of proposed method on CelebA-HQ, Places2 and Paris StreetView, respectively. By analyzing the images, proposed method has better artifact removal ability than MADF, TFill, AOT and W-Net, as shown in comparison results of third row in Fig. 6 and first row in Fig. 7. Then, inpainting results of proposed method have better ability to repair details, such as teeth and eyes of person in second row in Fig. 6. Above advantages are mainly due to feature filtering of MRFG and FMS modules. In addition, due to global context feature fusion capability based on decoder with FFC repair results of proposed method have better semantic consistency. For example, the word "cafe" on billboard in second row of Fig. 7.

GT Input MADF TFill W-Net AOT FCF Ours

Fig. 6 Comparison results of different methods on CelebA-HQ dataset. GT represents Ground Truth. Zoom in for more details

GT Input MADF TFill W-Net AOT FCF SCAT Ours

Fig. 7 Comparison results of different methods on Places2 dataset. GT represents Ground Truth. Zoom in for more details

GT Input MADF W-Net Ours

Fig. 8 Comparison results of different methods on Paris StreetView dataset. GT represents Ground Truth. Zoom in for more details

2.2.2 Quantitative comparisons

Quantitative comparison results of proposed method are shown in Tables 5, 6 and 7, respectively. According to the data in statistical table, comprehensive index of proposed method is better than other methods, and optimal rate of index reaches 66.67%. Specifically, proposed method achieves significant performance breakthrough in 0-40% damage ratio, while proposed method fails to achieve optimal repair index in 40-60% damage ratio. By analyzing above data, for large area is damaged, multi-stage inpainting method from coarse-to-fine repair coarse structure at first and then further repair according to coarse image, which is more conducive to filling large damaged area. However, single-stage inpainting method directly fills large empty areas when repairing, which is easily affected by damaged area. For example, FID index of the proposed method on Places2 dataset is only 46.95.

Table 5 Quantitative comparison results on CelebA-HQ dataset. ↑ and ↓ represent larger and smaller is better, respectively

	Mask ratio	MADF	TFill	W-Net	AOT	FCF	Ours
PSNR↑	1%-20%	34.52	33.74	34.54	31.69	34.75	36.53
	20%-40%	27.23	26.80	28.16	24.39	27.63	28.45
	40%-60%	22.35	22.72	23.91	19.98	23.36	23.16
	Average	28.03	27.75	28.87	25.35	28.58	29.38
SSIM↑	1%-20%	0.958	0.961	0.965	0.947	0.964	0.975
	20%-40%	0.874	0.881	0.895	0.836	0.885	0.907
	40%-60%	0.746	0.767	0.794	0.690	0.776	0.781
	Average	0.859	0.870	0.885	0.825	0.875	0.888
LPIPS↓	1%-20%	0.031	0.019	0.020	0.036	0.019	0.012
	20%-40%	0.085	0.060	0.058	0.103	0.056	0.048
	40%-60%	0.171	0.117	0.116	0.199	0.107	0.106
	Average	0.095	0.065	0.065	0.113	0.061	0.056
FID↓	1%-20%	4.44	2.59	3.07	5.59	2.96	1.73
	20%-40%	13.47	7.20	6.87	16.48	6.77	5.92
	40%-60%	36.45	13.79	11.64	44.22	10.13	15.32
	Average	18.12	7.86	7.19	22.10	6.62	7.66

Table 6 Quantitative comparison results on Places2 dataset. ↑ and ↓ represent larger and smaller is better, respectively

	Mask ratio	MADF	TFill	W-Net	AOT	FCF	SCAT	Ours
PSNR↑	1%-20%	30.94	30.67	31.15	30.28	31.11	28.74	32.39
	20%-40%	24.15	23.84	24.45	24.45	23.52	22.52	24.46
	40%-60%	20.21	20.14	20.54	20.06	19.24	19.06	19.85
	Average	25.10	24.88	25.38	24.93	24.62	23.44	25.57
SSIM↑	1%-20%	0.941	0.951	0.951	0.949	0.952	0.931	0.961
	20%-40%	0.823	0.844	0.845	0.851	0.839	0.802	0.857
	40%-60%	0.672	0.700	0.699	0.694	0.681	0.649	0.695
	Average	0.812	0.832	0.832	0.832	0.824	0.794	0.838
LPIPS↓	1%-20%	0.050	0.034	0.037	0.044	0.034	0.059	0.028
	20%-40%	0.132	0.111	0.116	0.101	0.109	0.149	0.106
	40%-60%	0.234	0.222	0.234	0.206	0.215	0.253	0.243
	Average	0.139	0.123	0.129	0.117	0.119	0.154	0.126
FID↓	1%-20%	7.49	6.24	7.57	7.71	6.04	9.93	5.00
	20%-40%	19.40	17.84	22.88	16.43	16.76	25.17	16.89
	40%-60%	34.40	34.05	49.00	33.54	29.33	43.18	46.95
	Average	20.43	19.38	26.48	19.23	17.38	26.09	22.95

Although the repair effect of proposed method is insufficient when large area damage, but still competitive with latest two-stage methods (e.g., for 40%-60% damage ratio, FID of W-Net is 49.00 on Places2 dataset, and proposed model has improvement about 4.2%) because of the enhancement

and filtering of intermediate features by MRFG and FMS modules, which is further verified and analyzed in ablation study.

Table 7 Quantitative comparison results on Paris StreetView dataset. ↑ and ↓ represent larger and smaller is better, respectively

	Method	1-20%	20-40%	40-60%	Average
PSNR↑	MADF	32.95	26.75	22.95	27.55
	W-Net	33.17	27.26	23.46	27.96
	Ours	34.67	27.46	22.72	28.28
SSIM↑	MADF	0.949	0.851	0.720	0.840
	W-Net	0.957	0.870	0.744	0.731
	Ours	0.967	0.879	0.731	0.859
LPIPS↓	MADF	0.040	0.106	0.192	0.112
	W-Net	0.033	0.094	0.190	0.106
	Ours	0.023	0.084	0.196	0.101
FID↓	MADF	19.11	41.50	64.92	41.84
	W-Net	18.52	39.25	69.30	42.36
	Ours	10.94	32.16	70.51	37.87

2.2.3 Experiment of large damage ratio

To further evaluate the performance of the proposed method in repairing images with large damage ratio, an experimental analysis was conducted using center square and corner square masks on the Paris StreetView dataset, and the results are shown in Table 8. For the center square damage, the proposed method holds a modest advantage in visual-related indicators such as LPIPS and FID, but there is a significant gap remains in pixel-level indicators such as PSNR and SSIM; for corner square damage, the inpainting performance of the proposed method is limited. As shown in Fig. 9, for center damage, the proposed method can repair relatively reasonable visual content, but the repaired result has inconsistencies in structure and pixels; for corner damage, the method yields less satisfactory visual outcomes with noticeable artifacts.

Table 8 Results of large damage repair on Paris StreetView dataset. ↑ and ↓ represent larger and smaller is better, respectively

Mask	Method	PSNR↑	SSIM↑	LPIPS↓	FID↓
Center square	MADF	24.98	0.846	0.115	36.11
	W-Net	25.53	0.854	0.121	39.58
	Ours	24.40	0.835	0.112	34.56
Corner square	MADF	26.54	0.874	0.090	30.52
	W-Net	27.11	0.877	0.099	33.65
	Ours	26.28	0.863	0.107	38.88

Ground Truth Input MADF W-Net Ours

Fig. 9 Comparison results of different methods on Paris StreetView dataset. GT represents Ground Truth. Zoom in for more details

2.3 Ablation study

To analyze the performance of proposed method, the ablation study section conducted decomposition experiments on Paris StreetView dataset and computational complexity analysis on Places2 dataset.

2.3.1 Decomposition experiment

As shown in Tables 9, a single codec was initially used as the baseline model. Subsequently, the FFC, MRFG and FMS modules were gradually introduced. By analyzing Table 9, thanks to the global contextual feature fusion capability of the FFC module, the overall performance of the network in the four indicators has significantly improved after introducing FFC into the decoder, especially at a damage ratio of 20% to 60%. Next, with the introduction of the MRFG module, the incorporation of multi-scale features led to improvement in visually relevant metric FID. However, other metrics slightly decreased due to the influence of redundant information in multi-scale features. Finally, the inpainting performance was further improved after introducing the FMS module to filter multi-scale features. A further analysis of the model settings in Table 9 shows that introducing FFC doubles the parameters, but the overall parameter count remains relatively small. In contrast, introducing both MRFG and FMS leads to a rapid increase in parameters due to the inclusion of

multi-receptive-field feature extraction and the nature of fully-connected feature selection. Future work may focus on optimizing the network through feature optimization and matching.

Table 9 The decomposition experiment of the proposed method on Paris StreetView dataset. ↑ and ↓ represent larger and smaller is better, respectively

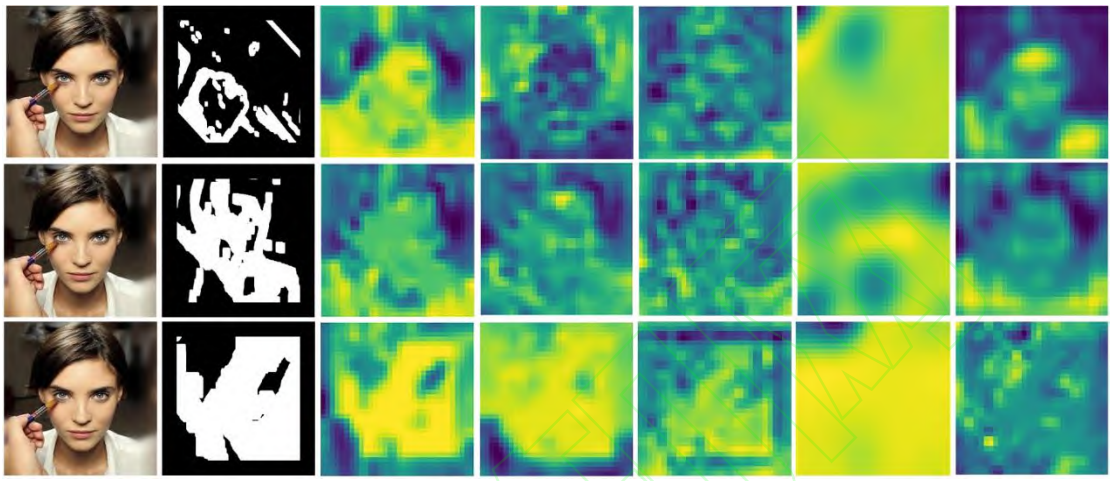
	Module Setting	1-20%	20-40%	40-60%	Average	Parameter
PSNR↑	(a) Base	33.61	26.36	21.54	27.17	5.52 M
	(b) Base+FCC	34.55	27.30	22.60	28.15	12.34 M
	(c) Base+FCC+MRFG	34.53	27.27	22.49	28.09	47.61 M
	(d) Base+FCC+MRFG+FMS	34.67	27.46	22.72	28.28	89.56 M
SSIM↑	(a) Base	0.9609	0.8587	0.6896	0.8364	5.52 M
	(b) Base+FCC	0.9668	0.8792	0.7348	0.8603	12.34 M
	(c) Base+FCC+MRFG	0.9666	0.8766	0.7238	0.8557	47.61 M
	(d) Base+FCC+MRFG+FMS	0.9672	0.8794	0.7306	0.8591	89.56 M
LPIPS↓	(a) Base	0.0297	0.1100	0.2509	0.1302	5.52 M
	(b) Base+FCC	0.0235	0.0861	0.2016	0.1037	12.34 M
	(c) Base+FCC+MRFG	0.0234	0.0874	0.2024	0.1044	47.61 M
	(d) Base+FCC+MRFG+FMS	0.0229	0.0838	0.1963	0.1010	89.56 M
FID↓	(a) Base	15.58	47.44	96.11	53.04	5.52 M
	(b) Base+FCC	12.01	34.79	75.37	40.73	12.34 M
	(c) Base+FCC+MRFG	11.38	33.68	72.67	39.25	47.61 M
	(d) Base+FCC+MRFG+FMS	10.94	32.16	70.51	37.87	89.56 M

Ground Truth Input (a) (b) (c) (d)

Fig. 10 Results of decomposition experiment for ablation study. (a) is pure convolutional codec. (b) add decoder with FFC. (c) add decoder with FFC and MRFG. (d) add decoder with FFC, MRFG, and FMS. Zoom in for more details

Furthermore, to provide a more intuitive demonstration the role of each module, a visual analysis of experimental results was conducted, as shown in Fig. 10. The single codec repair results have structural misalignment and blurring, as shown in Fig. 10(a), the network can obtain better detail repair results when repairing damaged regions with repeated textures after introduction of decoder with FFC module. (e.g., gray grid window in red-boxed area in first row of Fig. 10(b)), which verifies decoder with fast fourier convolution has better global contextual feature fusion

capability compared to pure convolutional decoder. Then, Fig. 10(c) shows that increase of multi-receptive field features strengthens constraint on repair results (e.g., lower half of window in red box in first row and iron grid in second row in Fig. 10(c)), but simple superposition of multi-receptive field features causes certain mixing effect (e.g., upper half of window in second row of Fig. 10(c)). Finally, Fig. 10(d) shows that FMS module eliminates mixing effect, and accuracy of repair results is further improved (e.g., the content in red box in curtain portion of first row and lower fence of second row in Fig. 10(f)).



(a) (b) (c) (d) (e) (f) (g)

Fig. 11 The intermediate feature maps of MRFG and FMS module. (a) and (b) are Ground Truth and Mask, respectively. (c), (d), (e) and (f) display the feature maps generated by MRFG, and receptive fields are 23×23, 46×46, 97×97, and 187×187, respectively. (g) is the feature map after weighting and filtering by FMS. Zoom in for more details

In addition, to analyze the mechanism of MRFG and FMS, this section further extracts and analyzes multi-receptive field feature maps generated by MRFG and feature maps weighted and filtered by FMS under different damage ratios. As shown in Fig. 11, damage ratios in first through third rows are 10-20%, 30-40%, and 50-60%, respectively. Analyzing features verifies that different receptive field features have varying sensitivities for different damage patterns. For example, high-dimensional features extracted by large receptive field in Fig. 11(f) are less sensitive to masks and details. However, small receptive field can retain detailed information of undamaged area, as shown in first and second rows of Fig. 11(c) and Fig. 11(d). Furthermore, by comparing first and second rows of Fig. 11(g) with other feature maps reveals that feature maps obtained by FMS after weighting and filtering multi-receptive field features are less affected by masks, which proves the effectiveness of FMS module.

2.3.2 Computational complexity analysis

A comparative experimental analysis of computational complexity was conducted using 256×256 images on an NVIDIA GeForce RTX 4090 GPU to evaluate the practical deployment value of the proposed model, with the results summarized in Table 10. Firstly, the proposed model has a moderate number of parameters (89.56 M) and storage requirements (342 MB). Secondly,

although the parameters is not significantly different from that of models such as FCF and TFill, the proposed model achieves a substantially higher inference speeds by deliberately avoiding complex structures during design. The processing time for a single image is approximately 0.019 seconds (52.63 FPS), which exceeds the standard 24 FPS used in films and offering potential for real-time inpainting applications. Thirdly, the maximum GPU memory consumption during model operation is about 3.25 GB, which means that the proposed model can be deployed on entry-level consumer GPUs such as NVIDIA GeForce RTX 4060. Finally, the proposed method has better inpainting performance although it is not as lightweight as SCAT. For example, on the Places2 dataset, the FID score is 22.95 for the proposed method and 26.09 for SCAT.

In summary, the proposed model has the potential for real-time inpainting and requires relatively low hardware resources. However, considerable improvements could still be made in terms of lightweighting and deployment on edge devices.

Table 10 Computational complexity analysis

Method	TFill	FCF	SCAT	Ours
Parameter	109.45 M	70.34 M	15.20 M	89.56 M
Infer.time(GPU)/per image	0.070 s	0.662 s	0.010 s	0.019 s
Frames/per second(FPS)	14.29	1.61	100.00	52.63
Model Size	417 MB	629 MB	58 MB	342 MB
Maximum allocated memory	0.74 GB	8.84 GB	0.12 GB	3.25 GB

2.4 Application

Dunhuang mural images were used to validate generalization ability of proposed model in this section, and results are shown in Fig. 12. More specifically, the simulated damage of mural image was performed, and proposed method can repair mural image in line with visual expectations when local information is missing.

Fig. 12 Example of mural image inpainting

3 Conclusion

This paper proposed an image inpainting network based on dynamic matching of multi-receptive fields and damaged patterns, which addresses the problem that current inpainting network have artifacts and semantic information confusion in repaired images due to ignoring adaptation between different receptive fields and image damaged patterns. Firstly, multi-scale receptive field features were extracted by parallel filter banks. Secondly, dynamic matching relationship between receptive field and damaged pattern was constructed by taking mask image as constraint condition. At the same time, decoder based on fast fourier convolution is designed to enhance global context feature fusion ability of pure convolution decoder. Extensive experiments show that proposed method can achieve SOTA results in small and medium damaged images. However, proposed model does not achieve SOTA results when large areas damaged. Since image inpainting task relies more on high-dimensional features of Fig. 11(f) when large area damaged. Although FMS of Fig. 11(g) over adds high-dimensional features and details to certain extent, and obvious mask traces are still visible due to heavily affected by mask. Therefore, how to fill in feature dimension when repair large area of continuous damaged is potential future optimization directions in future work.

References

[1] Wang H, Wang X, Liao R X, et al. Ship identification and tracking for navigable bridge based on improved YOLOv8 and ByteTrack[J]. Journal of Southeast University (Natural Science Edition), 2025, 55(5):1380-1387.

[2] Jia J W, Ni Y H, Mao J X, et al. Geometric parameter identification of bridge precast box girder sections based on deep learning and computer vision[J]. Journal of Southeast University (English Edition), 2025, 41(3):278-285.

[3] Xiang H Y, Zou Q, Nawaz M A, et al. Deep learning for image inpainting: A survey[J]. Pattern Recognition, 2023, 134:109046.

[4] Huang J B, Kang S B, Ahuja N, et al. Image completion using planar structure guidance[J]. ACM Transactions on Graphics, 2014, 33(4):1-10.

[5] Zhao X, Wang L M, Zhang Y F, et al. A review of convolutional neural networks in computer vision[J]. Artificial Intelligence Review, 2024, 57(4):99.

[6] Huo G Y, Sheng X L, Shi S X. Unsupervised underwater image enhancement with global background light intensity estimation and contrast adjustment[J]. Journal of Southeast University (Natural Science Edition), 2025, 55(1):297-305.

[7] Liu J L, Gong M G, Tang Z D, et al. Deep image inpainting with enhanced normalization and contextual

attention[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(10):6599-6614.

[8] Nazeri K, Ng E, Joseph T, et al. EdgeConnect: Structure guided image inpainting using edge prediction[C]//Proceedings of the IEEE International Conference on Computer Vision Workshop. Seoul, Korea, 2019:3265-3274.

[9] Zhang R S, Quan W Z, Zhang Y, et al. W-Net: Structure and texture interaction for image inpainting[J]. IEEE Transactions on Multimedia, 2023, 25:7299-7310.

[10] Li Y, Zhai J, Lu W, et al. Image inpainting with aggregated convolution progressive network[J]. IET Image Processing, 2025, 19(1):13318.

[11] Liu S, Chen J Y, Ding X L, et al. Progressive reverse attention network for image inpainting detection and localization[J]. Computer Vision and Image Understanding, 2025, 259:104407.

[12] Zheng C X, Cham T, Cai J F, et al. Bridging global context interactions for high-fidelity image completion[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA, 2022:11512-11522.

[13] Yu J H, Lin Z, Yang J M, et al. Generative image inpainting with contextual attention[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, 2018:5505-5514.

[14] Liu H Y, Jiang B, Song Y B, et al. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations[C]//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020:725-741.

[15] Liu G L, Reda F, Shih K, et al. Image inpainting for irregular holes using partial convolutions[C]//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018:85-100.

[16] Johnson J, Alahi A, Li F F. Perceptual losses for real-time style transfer and super-resolution[C]//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016:694-711.

[17] Gatys L, Ecker A, Bethge M. Image style transfer using convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA, 2016:2414-2423.

[18] Doersch C, Singh S, Gupta A, et al. What makes paris look like paris?[J]. Communications of the ACM, 2015, 58(12):103-110.

[19] Liu Z W, Luo P, Wang X G, et al. Deep learning face attributes in the wild[C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015:3730-3738.

[20] Zhou B L, Lapedriza A, Khosla A, et al. Places: A 10 million image database for scene recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(6):1452-1464.

[21] Quan W Z, Zhang R S, Zhang Y, et al. Image inpainting with local and global refinement[J]. IEEE Transactions

on Image Processing, 2022, 31:2405-2420.

[22] Wang Z, Bovik A, Sheikh H, et al. Image quality assessment: From error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4):600-612.

[23] Heusel M, Ramsauer H, Unterthiner T, et al. GANs trained by a two time-scale update rule converge to a local nash equilibrium[C]//Proceedings of the International Conference on Neural Information Processing Systems. Long Beach, CA, USA, 2017:6629-6640.

[24] Zhang R, Isola P, Efros A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, 2018:586-595.

[25] Zhu M Y, He D L, Li X, et al. Image inpainting by end-to-end cascaded refinement with mask awareness[J]. IEEE Transactions on Image Processing, 2021, 30:4855-4866.

[26] Zeng Y H, Fu J L, Chao H Y, et al. Aggregated contextual transformations for high-resolution image inpainting[J]. IEEE Transactions on Visualization and Computer Graphics, 2023, 29(07):3266-3280.

[27] Jain J, Zhou Y Q, Yu N, et al. Keys to better image inpainting: Structure and texture go hand in hand[C]//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Waikoloa, HI, USA, 2023:208-217.

[28] Zuo Z W, Zhao L, Li A L, et al. Generative image inpainting with segmentation confusion adversarial training and contrastive learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington DC, USA, 2023:3888-3896.