# Efficient Hierarchical Feature Collaboration Transformer for Image Inpainting

Dengyong Zhang, Nuo Fu, Xin Liao, Jiaxin Chen, Hengfu Yang, and Gaobo Yang

***Abstract*—Existing image inpainting methods face limitations in detail restoration. Although transformer-based models have made certain progress recently, the lack of hierarchical feature interaction and insufficient consideration of the importance of features at different network levels lead to semantic ambiguity in image reconstruction. To enhance the visual quality and accuracy of image inpainting, we adopt a multi-level feature fusion approach and propose a novel, efficient hierarchical feature collaboration transformer (HFCT). Our approach comprises two modules: dual stream gated feature fusion (DSGF) and region-separated attention module (RSAM), effectively capturing features at different levels of the network and enhancing inter-level information exchange. The DSGF module uses soft gating to fuse primary and advanced features, strengthening the connection from local to global consistency and reducing artifacts. The RSAM module resolves attention isolation issues in feature fusion through region-separated attention, strengthening the understanding of feature relationships, capturing more image semantics, and improving restoration accuracy. Extensive experiments on the Paris StreetView, CelebA-HQ, and Places2 benchmark datasets demonstrate that our proposed method achieves superior image inpainting quality compared to several state-of-the-art inpainting algorithms. Please refer to the project page: https://github.com/csfunuo/HFCT.**

***Index Terms*—Image inpainting, hierarchical collaboration, region-separated attention, transformer.**

## I. Introduction

IMAGE inpainting [1] originates from image processing and has become a research hotspot with the rapid development of deep learning. It aims to fill in damaged or missing parts of an image while enhancing the visual aesthetics of the restored image, making it an essential task in computer vision. It is widely applied in repairing damaged photographs, restoring historical document images, and aiding in criminal investigations, among other fields. The key challenge of image inpainting is generating convincing structures and authentic textures to seamlessly integrate completed pixels with existing portions, achieving an aesthetically pleasing result.

Earlier image inpainting methods, such as diffusion-based methods [2], [3], fill in missing regions by propagating the appearance of neighboring content but are prone to generating noticeable artifacts. Since they lack consideration for the global image structure, these methods are mainly suitable for filling small holes but fail to deal with large-scale damage. Patch-based methods [4], [5] address missing regions through the search and utilization of image patches that are most similar to the part to be repaired and are usually more effective for small missing holes or simple background filling. However, both methods rely on structural similarities between image pixels for restoration. They are often limited to local information based on pixel correlations, lacking the utilization of higher-level semantic information from the image. This leads to a lack of realism and semantic coherence in the restoration results. In recent years, learning-based methods have made remarkable advancements, largely due to the prowess of convolutional neural networks (CNNs) being harnessed in grasping the intricate semantics of images [6]–[8]. While CNNs feature spatially sparse connectivity and parameter sharing, making them computationally efficient, their fixed features limit performance in inpainting tasks [9], [10]. Firstly, the parameter sharing and static nature of CNNs restrict adaptability to different features, easily leading to image structure distortion and texture blur. Secondly, CNNs operating within local windows struggle to capture global image structure information, resulting in locally distorted restoration with inconsistent artifacts when dealing with large missing areas. Finally, the complexity and computational load of CNNs limit their practical application in high-resolution image inpainting tasks.

Recently, inspired by the outstanding performance of the transformer [11] in the field of natural language processing, researchers have explored applying the standard transformer and its variants to various computer vision tasks achieving remarkable results in image inpainting tasks as well. Compared to CNNs, transformer have a broader receptive field, can better capture long-range dependencies, and provide better high-resolution restoration results. Wan et al. [12] was the first to apply the transformer to image inpainting, proposing the ICT algorithm that aimed to minimize computational costs through a two-stage restoration process but still required lengthy training times. As a result, Yu et al. [13] improved the

This work was funded in part by the National Natural Science Foundation of China under Grant 62172059, 62272160, U22A2030, 61972142 and 62402062, in part by National Key R&D Program of China under Grant 2024YFF0618800 and 2022YFB3103500, in part by Scientific Research Fund of Hunan Provincial Education Department of China under Grant 22A0200, in part by Hunan Provincial Funds for Distinguished Young Scholars under Grant 2024JJ2025, and in part by Natural Science Foundation of Changsha City, China under Grant kq2402031. (Corresponding author: Xin Liao)

Dengyong Zhang, Nuo Fu and Jiaxin Chen are with the School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410004, China (e-mail: zhdy@csust.edu.cn; 22208051655@stu.csust.edu.cn; jxchen@csust.edu.cn; ).

Xin Liao and Gaobo Yang are with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China (e-mail: xinliao@hnu.edu.cn; yanggaobo@hnu.edu.cn).

Hengfu Yang is with the School of Computer Science, Hunan First Normal University, Changsha 410205, China (e-mail: hengfuyang@hnfnu.edu.cn).

first stage of the ICT algorithm, leveraging long-range correlation and global structure modeling capabilities to enhance inpainting performance. These two-stage image inpainting methods employ attention mechanisms to reconstruct image features coarse-to-fine, with the coarse network handling semantic restoration and the fine network reconstructing texture and details. However, these methods have inevitable errors, increased training duration, and computational complexity. Furthermore, solely relying on single-modal similarity metrics to compute image context for information transfer, while seemingly reasonable, overlooks the semantic coherence of the image and lacks effective extraction and consideration of structural information. Li et al. [14] proposed the MAT model, which employs multi-head contextual attention mechanisms and sliding windows to establish long-range dependency priors, aiming to achieve higher quality and resolution in the restored images. While these models can generate high-quality image inpainting, they are still slightly inferior in the detail inpainting part of complex images. These studies highlight the advantages of attention mechanism-based transformer in image inpainting. However, existing inpainting models lack sufficient interaction between features at different levels. As a result, they can't obtain multi-level semantic information. This leads to blurry semantic reconstruction of images, especially for eyes and nose areas of facial images. Moreover, current transformer-based image inpainting models do not fully account for the mutual influence of features. Therefore, there is inadequate feature weighting. Also, they fail to measure complex dependencies and correlations between image content comprehensively.

To address the shortcomings of existing transformer-based inpainting methods that have not fully considered the mutual influence of hierarchical features, we propose a novel, efficient hierarchical feature collaboration transformer for image inpainting. This model combines the global feature encoding and long-range dependency, capturing the strengths of the transformer while reducing the number of parameters and computational cost. By leveraging an encoder-decoder framework, the model can generate visually pleasing content even for complex images. To better utilize the different levels of features, the features generated by the network are classified into three levels: primary features, which retain more intricate details such as structural and textural information; intermediate features, which represent the top-level data extracted during the encoding phase and encompass abundant global context; and advanced features, which expand the visual receptive field during the inpainting process. The main contributions of the work are enumerated as follows:

1) We propose the efficient hierarchical feature collaboration transformer (HFCT). Unlike existing methods, it addresses the issue of insufficient hierarchical interaction in transformer models by integrating multi-level features effectively, ensuring smooth information flow within the model.
2) We introduce the dual stream gated feature fusion module (DSGF), which merges different levels of primary and advanced features. This allows the network to learn multi-level encoded and decoded features efficiently, improving consistency in inpainting and reducing blurring and artifacts around holes.
3) We present the region-separated attention module (RSAM), which captures the relationships between primary and intermediate features. Mask participation in learning can focus on different parts of the image to capture more semantic information, resulting in finer inpainting and ensuring that the reconstructed image has a more coherent structure and finer textures.
4) Through extensive validation on three datasets, our proposed HFCT model outperforms existing image inpainting methods and achieves state-of-the-art performance.

The rest of this paper is organized as follows. Section II reviews the related works on image inpainting. Section III presents the architecture of the proposed HFCT, including detailed descriptions of the Dual Stream Gated Feature Fusion (DSGF) module and the Region-Separated Attention Module (RSAM). Section IV discusses the experimental results. Finally, Section V concludes the paper and suggests directions for future work.

## II. Related Work

### A. Traditional Inpainting

Traditional image inpainting methods rely on correlations and structural similarities between image pixels to infer the inpainted image. Depending on the approach, these methods can be classified as diffusion-based methods [15], [16] and patch-based methods [17], [18]. The former fills in missing regions by propagating the appearance of neighboring content. However, due to the limitations of the neighboring content search mechanism, the inpainted areas can exhibit noticeable artifacts within the image when faced with large areas of damage. The latter fills the missing regions by searching for patches most similar to the part to be repaired. Although these methods can capture long-range information, they are commonly more effective in addressing small missing holes or simple replenishment of scenes due to their inability to capture high-level semantics. As a result, they cannot effectively repair images with complex patterns or generate new objects that are not present and do not exist in the part to be repaired.

### B. Deep Learning Based Inpainting

The limitations of traditional image inpainting methods have prompted researchers to explore more effective solutions. With the evolution of CNNs, deep learning-based methods [6], [19], [20] have been widely explored in the field of image inpainting. Pathak et al. [9] proposed a context encoder approach based on deep neural networks for inpainting, introducing the framework of generative adversarial networks (GAN) [21] and pioneering the development of encoder-decoder architecture and adversarial training for the task of image inpainting. The approach infers missing image information using the known content of the image regions by employing an unsupervised visual feature learning algorithm driven by context pixel predictions. However, it overlooks the global regions of the image, resulting in structural inconsistencies in the inpainted

image. Iizuka et al. [6] proposed a fully convolutional network model by introducing a series of dilated convolution layers to increase the receptive field, including global and local discriminators. The joint training of global and local discriminators as adversarial losses generate inpainted images with consistent global and local semantics. Subsequently, some researchers were guided with additional image information for image inpainting. Nazeri et al. [22] proposed a two-stage adversarial inpainting model called EdgeConnect, which fills holes with preconditioned information and introduces edge information to enhance the structure. Likewise, Xiong et al. [23] adopted a contour generator as a structural prior instead of an edge generator, designing a similar model. Some methods have incorporated style-specific features into the inpainting process, the style-guided dual-branch inpainting network [24] refine content and style separately, effectively preserving artistic value and coherence.

Yang et al. [25] proposed a multitask learning framework that generates fine structures by introducing structural embeddings. Peng et al. [20] proposed a conditional autoregressive network and a structural attention module to learn the distribution of structural features and capture distance relationships between structures. However, these methods do not simultaneously utilize both structural and textural features, resulting in structural and textural inconsistencies in the inpainted image. Drawing inspiration from this, Liu et al. [26] created a mutual encoder-decoder network that learns both structural and textural features separately. Distinct from the simple encoder-decoder architecture, Li et al. [7] gradually repaired the missing regions by designing the region recognition module and the feature inference module to work in an alternating cycle while continuously merging all the feature maps generated in the inference process. Although the above methods utilize shallow features, their choice of directly stacking and fusing features is too simplistic. It often leads to useful information being buried in a large amount of redundant data, causing effective details suppressed in shallow layers to go undetected. The image context is not fully utilized, resulting in insufficient connections from local features to overall consistency.

CNNs have demonstrated their advantages in learning image semantics in image inpainting. However, their sparse spatial connectivity and fixed features limit performance in restoration tasks [8], [10]. The parameters of the convolutional kernel of CNNs remain constant during the inference process, which may lead to identical processing of valid, invalid, and hybrid features, especially when processing complex images prone to structural distortions and texture blurring. In addition, CNNs operate only within a local window, which makes it difficult to capture remote structural information, leading to local distortions and incoherence, especially when dealing with extensive damage. Finally, although CNNs perform well in image inpainting tasks, their complexity and high computational effort limit the operation speed and practical applications.

### C. Vision Transformer

In recent years, as the potential of transformer [11] in the field of image processing has become more widely recognized [27], [28], researchers have begun to apply their broader receptive fields and excellent performance at high resolutions to the field of image inpainting. Wan et al. [12] first used transformer for image inpainting by proposing the ICT algorithm. ICT is a two-stage repair method, where the first stage is reconstructed by transformer to obtain a coarse inpainting result with diverse global structures, followed by refining details with CNNs. Although the method minimizes computational load, the training time is too long. Therefore, Yu et al. [13] introduced a bidirectional autoregressive transformer as a refinement to the first stage of ICT, which enhances the ability to restore diverse and coherent image structures by utilizing its excellent ability to model long correlations and global structures. In order to obtain restored images inpainting with higher quality and resolution, Zheng et al. [29] introduced a restricted CNNs header on ICT and designed a TFill model based on content-based inference to mitigate neighborhood effects. The method performs well on irregularly masked images but cannot understand and visualize advanced semantic content.

Similarly, Li et al. [14] proposed a mask-guided transformer for large-area defects inpainting model, which combines the advantages of transformer and convolution to process high-resolution images efficiently and multi-head contextual attention, but the semantic comprehension and inference speed need to be improved. Zheng et al. [30] utilized the transformer encoder architecture containing multiple weighted bi-directional self-attention modules to capture global contextual dependencies, which improves the inference speed. The method outperformed ICT, but the models trained could not be generalized to arbitrary images. Cao et al. [31] used the vision transformer (VIT) [28] architecture as the backbone of the Masked Encoder after recognizing the great potential of VIT in visual recognition and target detection tasks, utilizing learning informative a priori features from pre-trained features to feed CNNs-based repair models. The high computational cost of ViT, especially for high-resolution restoration tasks, poses a significant limitation. Chen et al. [32] proposed a spatially-activated shannel attention layer with a sandwich structure for enhanced self-attention.

These studies show the superiority of transformer based on the attention mechanism for image inpainting. However, since transformer requires a large amount of computer memory, although the above methods reduce the computational burden to some extent, there are still some limitations in practical applications. First of all, there are fewer interactions between different layers of features in the model, which cannot fully capture the semantic information of each layer, resulting in insufficiently detailed image inpainting in terms of image semantics. Furthermore, the methods above exhibit limitations in adequately considering the intricate interactions among distinct features. They may suffer from insufficient feature weighting, failing to distinguish the importance of different features accurately and thus failing to measure the complex dependencies and correlations among image contents comprehensively. To address these shortcomings, we propose a transformer model that optimizes the interaction between multi-level features for more precise and accurate image inpainting.

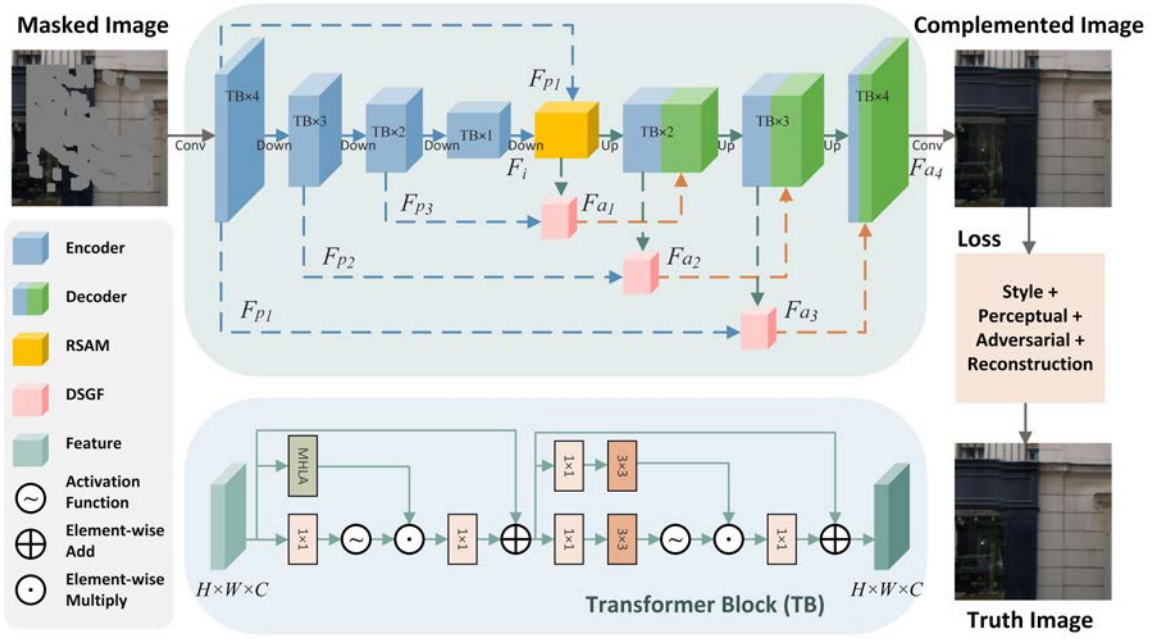


Fig. 1. The overview of the HFCT architecture takes a masked image as input and outputs an inpainted image. An encoder-decoder framework stacked by transformer blocks is used to learn the depth features of the image. We designed a dual stream gated feature fusion module (DSGF) and a region-separated attention module (RSAM) to jointly learn hierarchical features and fully exploit their contextual information.

## III. Approach

Our hierarchical feature collaboration transformer achieves high-quality image inpainting through multi-level feature extraction, fine fusion, and progressive decoding. This design adopts a hierarchical feature fusion strategy, where the dual stream gated feature fusion module (DSGF) integrates multi-level features to address the lack of effective information utilization during inpainting. This enhances the interaction of full-text information and ensures the coherent flow of effective information within the model. Additionally, the region-separated attention module (RSAM) captures the relationship between primary and intermediate features with the help of masks and finely separating regions to capture more image semantics. This approach effectively extracts long-range dependencies.

### A. Hierarchical Feature Collaboration Structure

Considering the advantages of transformer in encoding global features and capturing long-range dependencies, we constructed an inpainting network based on multiple transformer blocks. The design employs a hierarchical feature fusion strategy aimed at generating more detailed images through fusing features from different levels and scales, as well as subregional fusion involving masks. An overview of the proposed method is shown on Fig.1.

Given the input masked image $I_m \in \mathbb{R}^{H\times W\times 4}$ ($H$ for image height, $W$ for width), the mask occupies 1 channel while the image itself has 3 channels. Through a convolutional mapping, the image is initially transformed into a feature with $C$ channels ($C$ for the number of channels), resulting in $E_{in} \in \mathbb{R}^{H\times W\times C}$. Then, it is fed into a four-stage encoder. In each encoding stage, a feature is generated from the given input. Specifically, the features extracted by the encoders are denoted as $E_i \in \mathbb{R}^{\frac{H}{2^{i-1}}\times\frac{W}{2^{i-1}}\times C\times 2^{i-1}}$, with $i$ (ranging from 1 to 4) representing the number of encoding layers. Features from the first three layers of the encoder are defined as primary features $F_p$, as they retain more detailed information, such as structural and textural information. Downsampling is performed through convolution between stages, enabling the network to progressively capture different scales of information, facilitating better understanding in later stages. The final stage of the encoder outputs a feature map referred to as intermediate features $F_i$, which is the highest level feature extracted by the network during the encoding process and contains rich global context information.

The RSAM module is set in the middle of the network and receives the primary features $F_{p_1}$ from the first layer of the encoder and intermediate features $F_i$, as well as introducing masks. The primary features from the first layer are inherently closer to the original input image data, making them more suitable for assisting the inpainting process. Such selection optimizes the ability of the module to accurately capture and preserve critical image features during the inpainting process. Through region attention computation, we can achieve fine-grained fusion, flexibly focusing on different parts of the image, integrating feature information, and enhancing the understanding of image details and context. Subsequently, the feature generated by the RSAM module is sent to the decoder for decoding. The decoder section consists of three stages, each producing different feature maps that are gradually upsampled to restore high resolution, progressively restoring the details and structure of the image to achieve a finer inpainting result. The features extracted by the decoders are denoted as $D_i \in \mathbb{R}^{\frac{H}{2^{3-i}}\times\frac{W}{2^{3-i}}\times C\times 2^{3-i}}$, where $i$ ranges from

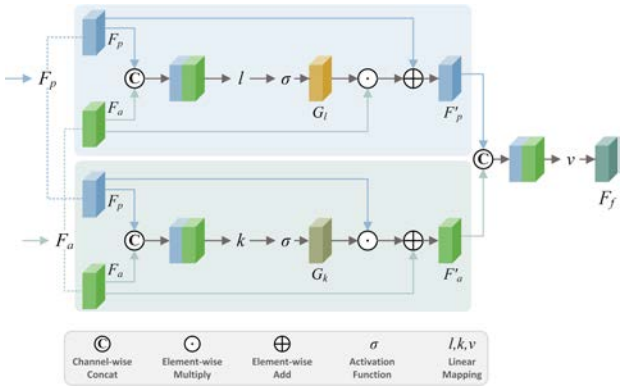


Fig. 2. Overview of DSGF. The module fuses primary and advanced features through soft gating to achieve tight integration of feature information at different levels.

1 to 3. We refer to the features generated by the RSAM and the decoder as advanced features $F_a$. These features extend the visual receptive field during the inpainting process, allowing for capturing broader image context information. By connecting the primary features $F_p$ and advanced features $F_a$ via the DSGF, we achieve the fusion of features from different levels, improving the consistency and accuracy of the inpainting process. Finally, a convolution layer is utilized to restore the image features to the original image size, obtaining $I_{out} \in \mathbb{R}^{H\times W\times 3}$.

Our basic architecture refers to the U-Net [33], employing a four-stage encoder and a three-stage decoder, each stage composed of different numbers of transformer blocks stacked in sequence, with the numbers being (4, 3, 2, 1, 2, 3, 4). The design of our transformer blocks references the encoder block in the vanilla transformer [11], each consisting of two parts: an efficient multi-head linear attention (MHLA) and a gated feed-forward network, both integrated through residual connections. The network achieves multi-scale feature extraction and restoration through progressive downsampling in the four encoder stages and upsampling in the three decoder stages. To accommodate the different stages of the encoders and decoders, the multi-head attention parameters of the transformer blocks vary as (1, 2, 4, 8, 4, 2, 1). This design enables the model to capture detailed information at different scales.

### B. Dual Stream Gated Feature Fusion

The DSGF is proposed to better integrate the hierarchical features of the whole network. As depicted in Fig.2, it can explore different feature levels and capture the primary features $F_p$ output from the encoder and the multi-level advanced features $F_a$ to integrate feature information from different hierarchical levels better and improve the consistency of the restoration effect. The DSGF employs a soft-gating fusion to dynamically adjust the fusion ratio of the primary and advanced features using the learning parameters to realize the information interaction. This approach enables the network to learn differentiated multimodal coded features compactly, ensuring that the network can obtain deep semantic features and avoid the loss of texture detail information in shallow features. Specifically, soft-gated $G_k$ is utilized to control a certain degree of refinement of the primary feature encoding information to construct advanced features guided by primary features. The soft gating control can be defined as:

$$G_k = \sigma\left(k\left(\left[F_p, F_a\right]\right)\right) \tag{1}$$

where $k(\cdot)$ is the convolution layer with kernel size 3, and $\sigma(\cdot)$ is the activation function Sigmoid. With soft-gated $G_k$, $F_p$ can be dynamically merged into $F_a$ in the following:

$$F_a^{'} = \lambda\left(G_k \odot F_p\right) \oplus F_a \tag{2}$$

where $\lambda$ is a learnable training parameter, $\odot$ denote elemental multiplication and $\oplus$ denote elemental addition.

Similarly, the primary features guided by advanced features $F_p'$ can be defined as:

$$G_l = \sigma\left(l\left(\left[F_p, F_a\right]\right)\right) \tag{3}$$

$$F_p^{'} = \mu\left(G_l \odot F_a\right) \oplus F_p \tag{4}$$

where $l(\cdot)$ is the convolution layer with kernel size 3 and $\mu$ is a learnable training parameter. Next, $F_p'$ and $F_a'$ are spliced together by channel. Then, a convolution layer $\upsilon(\cdot)$ with a kernel size of 1 is used to yield the combined feature representation $F_f$ for subsequent feeding to the corresponding decoder for further learning. Such a design effectively facilitates the interaction and integration between semantic and texture information so that the subsequent modules can learn richer hierarchical features.

$$F_f = \upsilon\left(\left[F_p^{'}, F_a^{'}\right]\right) \tag{5}$$

### C. Region-separated Attention Module

In addition to the attention in the transformer block, we designed a fine fusion module RSAM composed of regionally separated attention in the middle of the network to get more detailed restore results. The structure of RSAM is shown on Fig.3, which replicates the effective information from the primary feature $F_p$ and the intermediate feature $F_i$ and explicitly separates the masked region from the unmasked region for attention calculation. By separating the attention computation, RSAM can accurately deliver the global context information to the to-be-repaired region, thus providing richer details for the part to be repaired and making the result more accurate and fine.

Firstly, the input intermediate feature $F_i \in \mathbb{R}^{B\times C\times N}$, where $B$ denotes the batch size, $C$ denotes the number of channels, assuming $N = H \cdot W$, $H$ denotes the height of the image, $W$ denotes the width of the image, $N$ denotes the product of spatial dimensions. Two kinds of embeddings are obtained by $q(\cdot)$ and $k(\cdot)$ transformations: the query $Q \in \mathbb{R}^{B\times C\times N}$ and the key $K \in \mathbb{R}^{B\times C\times N}$, then calculate the attention weight $A$:

$$A = q\left(F_i\right)^T k\left(F_i\right) \tag{6}$$

where $q(\cdot)$ and $k(\cdot)$ are the mapping functions implemented by a convolutional layer with kernel size 1.

During the inpainting process, the features generated from the real regions and the restored masked regions differ, and

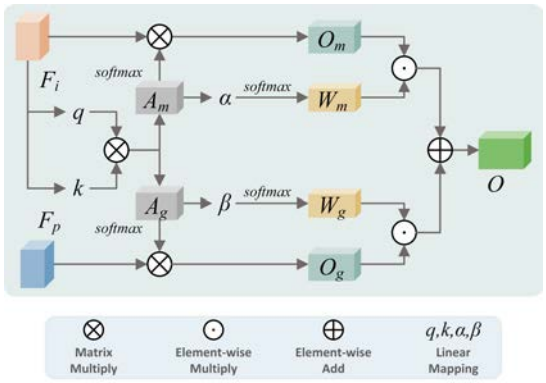


Fig. 3. Overview of RSAM. It accurately conveys global contextual information to the region to be repaired by combining primary and intermediate features and introducing masks for region-separated attention computation.

thus, attention isolation occurs during the feature fusion process. This means that when calculating the attention weights for the real regions, the network tends to focus more on the other real regions and ignore the masked regions and vice versa. In this case, with the use of complete attention at the deepest level of the network for feature fusion, the effect can not reach the expected level.

To solve this problem, the attention distribution is adjusted according to the binary mask $M$ by dividing the attention matrix $A$ into two parts: $A_g$ denotes the similarity associated with the real region, and $A_m$ denotes the similarity associated with the generated masked region. In this way, the attention weight has different values for different positions. The computations are defined as follows:

$$A_g = \sum_{i,j} M_{ij} A_{ij} \tag{7}$$

$$A_m = \sum_{i,j} (1 - M_{ij}) A_{ij} \tag{8}$$

Where $M_{ij} = 1$ indicates real regions and $M_{ij} = 0$ denotes masked regions.

After the two parts are normalized by softmax, the primary feature $F_p$ and intermediate feature $F_i$ are added to obtain the attention results of each region, to enhance the understanding and utilization of features at each level, better maintain the coherence of structure and consistency of texture, and thus improve the accuracy of the generated image.

$$O_g = softmax\left(A_g\right) F_p \tag{9}$$

$$O_m = softmax\left(A_m\right) F_i \tag{10}$$

Next, $O_g$ and $O_m$ are dynamically fused using the learnable parameters $\alpha$ and $\beta$. Instead of combining $O_g$ and $O_m$ by a fixed ratio, the fusion process learns the weight mapping based on the attention scores of each position. The fusion weights $W_g$ and $W_m$ are calculated as follows:

$$[W_g, W_m] = softmax\left(\left[\alpha\left(A_g\right), \beta\left(A_m\right)\right]\right) \tag{11}$$

where $\alpha$ and $\beta$ are 1×1 filters used to adjust the weights $W_g$ and $W_m$, softmax normalization is applied to ensure that $W_g + W_m = 1$ at each spatial location.

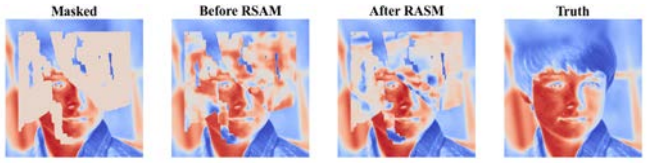


Fig. 4. Heatmaps of RSAM effects on inpainting: Masked, Before RSAM, After RSAM and Truth.

Finally, the fused attention result $O$ is obtained by the following equation:

$$O = W_g \cdot O_g + W_m \cdot O_m \tag{12}$$

To illustrate the role of RSAM in restoration, Fig.4 presents a visual comparison. The figure includes the masked input showing the inpainting region, the result before RSAM with uneven attention distribution and suboptimal feature recovery, the result after RSAM with improved attention distribution and restored feature details, and the ground truth as the reference. Comparing the results before and after RSAM reveals that RSAM directs more attention to the masked region while maintaining contextual coherence. The result after RSAM shows that the right eye and left ear of the covered face are reconstructed with clearer details and more distinct color blocks compared to the result before RSAM. This demonstrates that RSAM enables the network to prioritize global contextual information in the masked area, alleviate attention segregation, and achieve more accurate and natural reconstruction of missing details.

### D. Loss Functions

In order to obtain high-quality inpainting results, our HFCT is optimized by a joint loss $L_{joint}$ consisting of four terms: reconstruction loss $L_1$, perceptual loss $L_{prec}$ [34], style loss $L_{style}$ [35], and adversarial loss $L_{adv}$ [21]. Each loss function will be described in detail below.

**Reconstruction loss** The reconstruction loss is the value of the $L_1-$distance between the inpainting image $I_{out}$ and the real image $I_g$:

$$L_1 = \left\| I_{out} - I_g \right\|_1 \tag{13}$$

**Perceptual loss** The perceptual loss measures the feature map between the inpainting image $I_{out}$ and the real image $I_g$, defined as:

$$L_{prec} = \mathbb{E}\left[ \sum_i \left\| \phi_i\left(I_{out}\right) - \phi_i\left(I_g\right) \right\|_1 \right] \tag{14}$$

$\phi_i$ denotes the feature map of layer i of VGG-19 [36] pre-trained on ImageNet.

**Style loss** The style loss is defined similarly to perceptual loss:

$$L_{style} = \mathbb{E}\left[ \sum_i \left\| \varphi_i\left(I_{out}\right) - \varphi_i\left(I_g\right) \right\|_1 \right] \tag{15}$$

where $\varphi_i = \phi_i^T \phi_i$ represents the Gram matrix constructed by the active mapping $\phi_i$.

TABLE I
QUANTITATIVE COMPARISON ON THE PARIS STREETVIEW. ↓ MEANS LOWER IS BETTER, ↑ MEANS HIGHER IS BETTER.

| Metrics | Mask Ratio | GC [10] | RFR [7] | CTSDG [38] | HAN [39] | SPA [40] | Ours |
|---|---|---|---|---|---|---|---|
| LPIPS↓ | 10%-20% | 0.035 | 0.041 | 0.154 | 0.024 | 0.024 | **0.022** |
| | 20%-30% | 0.059 | 0.059 | 0.170 | 0.043 | 0.043 | **0.041** |
| | 30%-40% | 0.086 | 0.076 | 0.192 | 0.066 | 0.067 | **0.064** |
| | 40%-50% | 0.120 | 0.101 | 0.223 | 0.097 | 0.099 | **0.094** |
| PSNR↑ | 10%-20% | 31.204 | 30.826 | 31.841 | 32.691 | 32.778 | **33.022** |
| | 20%-30% | 28.092 | 28.501 | 28.806 | 29.745 | 29.758 | **29.961** |
| | 30%-40% | 26.237 | 27.152 | 26.562 | 27.521 | 27.502 | **27.622** |
| | 40%-50% | 24.390 | 25.752 | 24.525 | 25.534 | 25.585 | **25.693** |
| SSIM↑ | 10%-20% | 0.941 | 0.932 | 0.945 | 0.952 | 0.952 | **0.954** |
| | 20%-30% | 0.896 | 0.892 | 0.903 | 0.915 | 0.914 | **0.916** |
| | 30%-40% | 0.847 | 0.852 | 0.848 | 0.869 | 0.868 | **0.870** |
| | 40%-50% | 0.786 | 0.804 | 0.781 | 0.810 | 0.810 | **0.814** |
| FID↓ | 10%-20% | 23.024 | 27.854 | 18.736 | 15.360 | 14.189 | **13.552** |
| | 20%-30% | 37.525 | 40.016 | 35.312 | 25.973 | 24.447 | **23.598** |
| | 30%-40% | 50.848 | 50.445 | 53.190 | 37.511 | 37.138 | **34.644** |
| | 40%-50% | 64.546 | 63.254 | 76.290 | 49.529 | 47.762 | **47.388** |

**Adversarial loss** The adversarial loss is used to ensure the visual realism of the inpainting image as well as the consistency of the texture and structure, defined as:

$$L_{adv} = \mathbb{E}_{I_g}\left[logD\left(I_g\right)\right] + \mathbb{E}_{I_{out}} log\left[1 - D\left(I_{out}\right)\right] \quad (16)$$

where $D$ is the patch GAN discriminator [37] with spectral normalization.

**Joint loss** The joint loss used to train our HFCT can be defined as:

$$L_{joint} = \lambda_1 L_1 + \lambda_p L_{prec} + \lambda_s L_{style} + \lambda_a L_{adv} \quad (17)$$

where $\lambda_1$, $\lambda_p$, $\lambda_s$ and $\lambda_a$ are the corresponding weight parameters, which are set as $\lambda_1 = 1$, $\lambda_p = 1$, $\lambda_s = 250$ and $\lambda_a = 0.1$ in the experiment according to experience.

## IV. EXPERIMENTS

In this section, we present the experimental results to validate the superior performance of HFCT. For a full presentation, we first introduce the three publicly available datasets and elaborate on the detailed experimental setup. We then provide quantitative results on three datasets compared to five SOTA approaches. We have also attached the relevant qualitative results for a more intuitive presentation. Subsequently, we conduct an efficiency comparison experiment. Finally, we conducted extensive ablation experiments.

### A. Datasets

We evaluated our HFCT on three public datasets to verify the scientific validity and effectiveness of the current algorithms: Paris StreetView [9], CelebA-HQ [41], and Places2 [42]. The Paris StreetView dataset is architectural images from the street, containing 14,900 training images and 100 test images. For Paris StreetView, the standard configuration is followed to split the data for training and testing. The CelebA-HQ dataset contains 30,000 high-quality face images, using the first 28000 for training and the remaining 2000 for testing. The Places2 dataset contains more than 8 million images taken in more than 365 scenes, providing diverse scenes. During the experiment, 1000 images are selected from the test set for network performance testing. All images used for the experiments are resized to 256×256 pixels. In evaluating the network performance, the Irregular masks dataset proposed by Liu et al. [8] is used to simulate the damaged areas for quantitative evaluation. It contains 12000 irregular masks, where the area of each mask is 0% to 60% of the total image size and is divided into six intervals based on the mask area. Each mask interval contains 2000 irregular mask images.

### B. Experimental Settings

Our HFCT is implemented based on Pytorch, training and testing on two NVIDIA 2080Ti GPUs (12 GB). The AdamW optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.9$ was used to train the model, batch size was 4, the learning rate was set to $10^{-4}$, and later it was adjusted to $10^{-5}$ to fine-tune the model. Specifically, on CelebA-HQ and Paris StreetView, we trained $600,000$ iterations and then fine-tuned $300,000$ iterations. On the Places2 dataset, we trained about $1000,000$ iterations and fine-tuned $600,000$ iterations. All experimental results are output directly from the trained model without any post-processing.

**Evaluation metrics** In evaluating our network, a quantitative assessment of the model performance is followed using a set of evaluation metrics: peak signal-to-noise ratio (PSNR), structural similarity (SSIM), Fréchet Inception Distance (FID), and learned perceptual image patch similarity (LPIPS). PSNR measures the pixel-level similarity of the two images, and SSIM measures the brightness of the two images, contrast, and structure. FID [43] serves as a quantitative measure widely adopted in image generation to evaluate the image distribution between the repair result and the real image. LPIPS [44] is a learned perceptual image similarity metric that detects complex distortions consistent with subjective human perception.

**Baselines** Our network is compared with five models as baseline. These models include GC [10], RFR [7], CTSDG [38], HAN [39] and SPA [40]. GC [10] is a two-stage CNNs-based inpainting model that utilizes an adaptive feature selection gating mechanism and contextual attention to generate coherent images. RFR [7] is a recurrent inpainting

TABLE II
QUANTITATIVE COMPARISON ON THE CELEBA-HQ. ↓ MEANS LOWER IS BETTER, ↑ MEANS HIGHER IS BETTER.

| Metrics | Mask Ratio | GC [10] | RFR [7] | CTSDG [38] | HAN [39] | SPA [40] | Ours |
|---|---|---|---|---|---|---|---|
| LPIPS↓ | 10%-20% | 0.025 | 0.050 | 0.023 | 0.017 | 0.015 | **0.013** |
| | 20%-30% | 0.041 | 0.065 | 0.048 | 0.031 | 0.028 | **0.024** |
| | 30%-40% | 0.058 | 0.086 | 0.083 | 0.049 | 0.044 | **0.038** |
| | 40%-50% | 0.081 | 0.116 | 0.126 | 0.071 | 0.065 | **0.056** |
| PSNR↑ | 10%-20% | 32.133 | 30.642 | 32.788 | 33.074 | 33.485 | **34.334** |
| | 20%-30% | 29.199 | 28.180 | 28.927 | 29.846 | 30.262 | **30.944** |
| | 30%-40% | 26.968 | 26.047 | 26.038 | 27.316 | 27.713 | **28.376** |
| | 40%-50% | 25.076 | 24.062 | 23.774 | 25.224 | 25.529 | **26.206** |
| SSIM↑ | 10%-20% | 0.954 | 0.935 | 0.961 | 0.961 | 0.964 | **0.968** |
| | 20%-30% | 0.922 | 0.899 | 0.922 | 0.929 | 0.934 | **0.940** |
| | 30%-40% | 0.885 | 0.855 | 0.872 | 0.891 | 0.897 | **0.906** |
| | 40%-50% | 0.842 | 0.802 | 0.814 | 0.845 | 0.853 | **0.865** |
| FID↓ | 10%-20% | 3.833 | 11.257 | 3.673 | 2.298 | 2.079 | **1.717** |
| | 20%-30% | 5.253 | 12.744 | 8.884 | 3.862 | 3.436 | **3.055** |
| | 30%-40% | 7.290 | 16.810 | 19.770 | 5.616 | 5.005 | **4.454** |
| | 40%-50% | 9.916 | 24.320 | 39.208 | 7.703 | 6.765 | **6.027** |

TABLE III
QUANTITATIVE COMPARISON ON THE PLACES2. ↓ MEANS LOWER IS BETTER, ↑ MEANS HIGHER IS BETTER.

| Metrics | Mask Ratio | GC [10] | RFR [7] | CTSDG [38] | HAN [39] | SPA [40] | Ours |
|---|---|---|---|---|---|---|---|
| LPIPS↓ | 10%-20% | 0.056 | 0.042 | 0.043 | 0.037 | 0.033 | **0.032** |
| | 20%-30% | 0.094 | 0.076 | 0.084 | 0.066 | 0.061 | **0.058** |
| | 30%-40% | 0.135 | 0.116 | 0.133 | 0.102 | 0.096 | **0.091** |
| | 40%-50% | 0.185 | 0.163 | 0.190 | 0.149 | 0.139 | **0.133** |
| PSNR↑ | 10%-20% | 27.381 | 28.018 | 28.695 | 28.128 | 28.691 | **28.934** |
| | 20%-30% | 24.237 | 24.802 | 25.321 | 25.142 | 25.593 | **25.761** |
| | 30%-40% | 22.091 | 22.509 | 22.943 | 22.915 | 23.310 | **23.456** |
| | 40%-50% | 20.396 | 20.667 | 21.067 | 21.048 | 21.373 | **21.505** |
| SSIM↑ | 10%-20% | 0.922 | 0.928 | 0.937 | 0.930 | 0.936 | **0.938** |
| | 20%-30% | 0.861 | 0.869 | 0.880 | 0.875 | 0.883 | **0.886** |
| | 30%-40% | 0.796 | 0.801 | 0.813 | 0.811 | 0.821 | **0.826** |
| | 40%-50% | 0.723 | 0.725 | 0.737 | 0.737 | 0.749 | **0.754** |
| FID↓ | 10%-20% | 19.730 | 15.689 | 16.091 | 12.866 | 10.925 | **10.216** |
| | 20%-30% | 31.235 | 26.737 | 29.631 | 20.595 | 18.349 | **17.350** |
| | 30%-40% | 42.022 | 38.505 | 44.551 | 29.237 | 26.009 | **24.585** |
| | 40%-50% | 55.493 | 52.077 | 64.223 | 40.221 | 35.791 | **33.870** |

model that exploits knowledge-consistent attention to complete images and ensures consistency progressively. CTSDG [38] is a CNNs-based image inpainting that uses image texture and structure information to assist in obtaining detailed complementary images. HAN [39] is an hourglass attention network for image completion, leveraging a novel distance prior to enhancing contextual information synthesis for high-quality results. Spa [40] is a modified transformer block integrated into the U-Net framework for image inpainting, which achieves superior results with reduced computational requirements.

### C. Quantitative Comparison

The quantitative results are shown in Tables I, II, and III, which demonstrate the quantitative evaluation results of the Paris StreetView, CelebA-HQ, and Places2 datasets on irregularly masked datasets, respectively. HFCT performs better on all datasets and masking rates than the five state-of-the-art methods. On the CelebA-HQ dataset, the model performs well at a 10%-20% masking rate, and the PSNR metric improves by 0.849 compared to SPA [40], which shows better image inpainting capability. This is due to the multi-level feature extraction and level-by-level decoding strategies, which enable the network to better capture and recover image details at different scales and improve image quality. While at a 40%-50% masking rate, the model improves by 0.012 in the SSIM metric and reduces 0.025 and 3.889 in LPIPS and FID metrics, respectively, relative to SPA [40].

Similarly, the inpainting results based on Paris StreetView and Places2 show that the proposed method presents significant superiority in LPIPS, PSNR, SSIM, and FID metrics indicators. It is attributed to the design of the DSGF module in the network, which effectively fuses the primary features and advanced features to harmonize the structural and textural features extracted across the multi-level feature stage and improves the consistency and accuracy of the network in the image inpainting process. These results demonstrate that HFCT as an image inpainting method can provide stable restoration results with good generalization ability. At a large mask rate of 40%-50%, the method shows excellent performance compared to other models. This advantage comes from the introduction of regional attention computation, which realizes fine fusion in the middle of the network, enabling the network to better reconstruct and infer the details and contextual information in the image, thus improving the consistency of image inpainting results. The significant improvement in the FID metric is particularly noteworthy, which indicates that the HFCT can more accurately infer missing structures and textures in images.

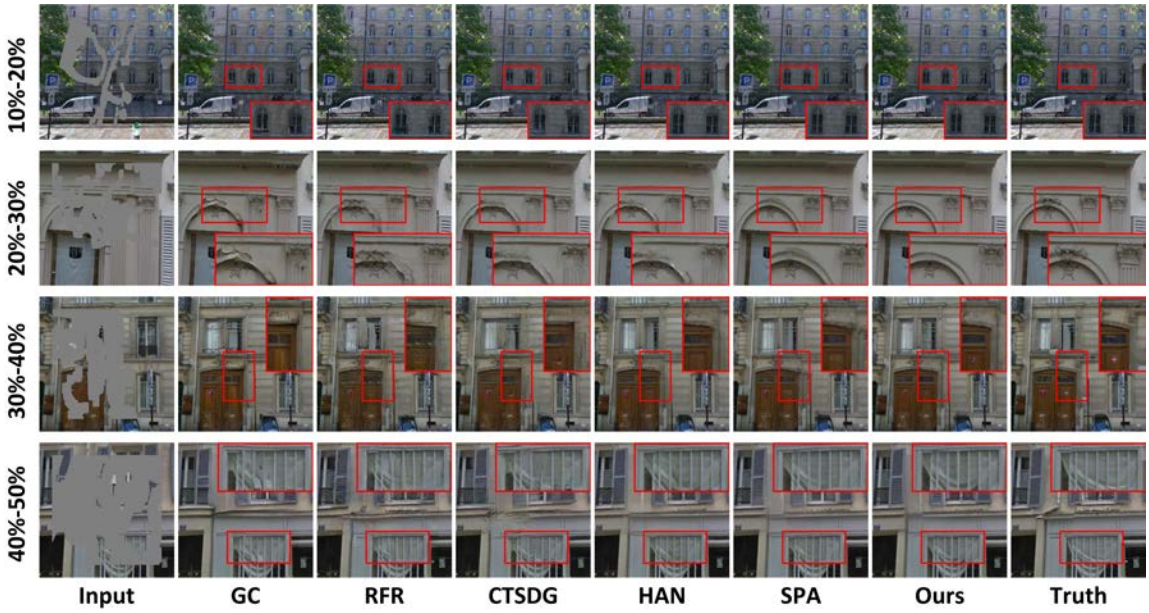


Fig. 5. Qualitative comparison with GC [10], RFR [7], CTSDG [38], HAN [39], and SPA [40] on the Paris StreetView dataset. The first and second rows exhibit mask rates of 10%-20% and 20%-30% respectively, while the remaining rows are generated with mask rates of 30%-40% and 40%-50%. Zooming in on the red box shows a better view of the local details.

These innovative designs enhance the restoration of network and generalization performance, especially at large mask rates.

### *D. Qualitative Comparisons*

Fig.5 illustrates the visual results from the Paris StreetView dataset. These examples depict the overall quality of the inpainting results generated at four masking rates, with local details demonstrated by zooming in on the restoration results through the red boxes. As the masking ratio increases, the images experience more extensive damage, making the inpainting process more challenging. In images with complex patterns, GC [10], RFR [7], and CTSDG [38] are prone to noticeable blurring, distorted edges, and disordered structures, such as the windowsill in row 1 and the curtains in row 4. While the results generated by HAN [39] and SPA [40] are generally satisfactory, some instances show incomplete or unclear inpainting, such as the arch in row 2 and the door logo in the row 3. In contrast, our model, HFCT, consistently delivers clear, smooth lines, achieving accurate semantic inpainting with refined texture details.

Fig.6 shows that HFCT better recovers finer-grained details, especially eye features, for face image inpainting compared to other methods. When the masking rate is low (rows 1 and 2), the differences in inpainting quality among almost all models are minimal. However, when it comes to a high masking rate (row 3), the detailed textures of the images supplemented by GC [10] and SPA [40] look good but are prone to semantic inconsistencies. Even though HAN [39] can make the pupil colour of the restored eye close to the same as that of the other real eye, it still needs improvement. In contrast, HFCT not only achieves the same pupil colour for both eyes but also has eye highlights, and the details of the ear part of the restoration are closer to the real situation. In the case of very large breakage, although GC [10] and HAN [39] can accomplish the basic semantic filling, the filling position is easily blurred, and the restoration results of the other methods even introduce many artifacts and obvious structural distortions. It is clear from row 4 that even under extreme damage(with a mask rate of 40%-50%), HFCT can still produce a more natural and detailed high-fidelity face image. The visualization results on the Places2 dataset are similar, as shown in Fig.7. The model demonstrates the capability of restoring intricate textures and details, which can be observed from the distinct boat structure in row 3. Furthermore, HFCT effectively manages complex scenes, like the building beam in row 4, revealing its strong inpainting abilities across diverse and challenging situations.

### *E. Efficiency Comparison*

In Table IV, we present an analysis of the parameter count and computational complexity of HFCT in comparison with five other models. Apart from GC [10] and SPA [40], HFCT maintains a relatively low parameter count and computational cost. Notably, while our model shows a modest 0.63% increase in parameters over GC [10], it achieves an approximately 40.3% reduction in computational cost. Although our parameter count is slightly higher than that of SPA [40] , this trade-off results in a marked improvement in inpainting quality. To further illustrate the balance between efficiency and performance, we conducted a detailed quantitative comparison with GC [10] and SPA [40], as shown in Tables I, II, and III. Across multiple datasets and evaluation metrics, our model consistently outperforms these baselines, demonstrating that even with a slight increase in parameters, HFCT offers significant performance gains on challenging inpainting tasks. These improvements can be attributed to our carefully designed

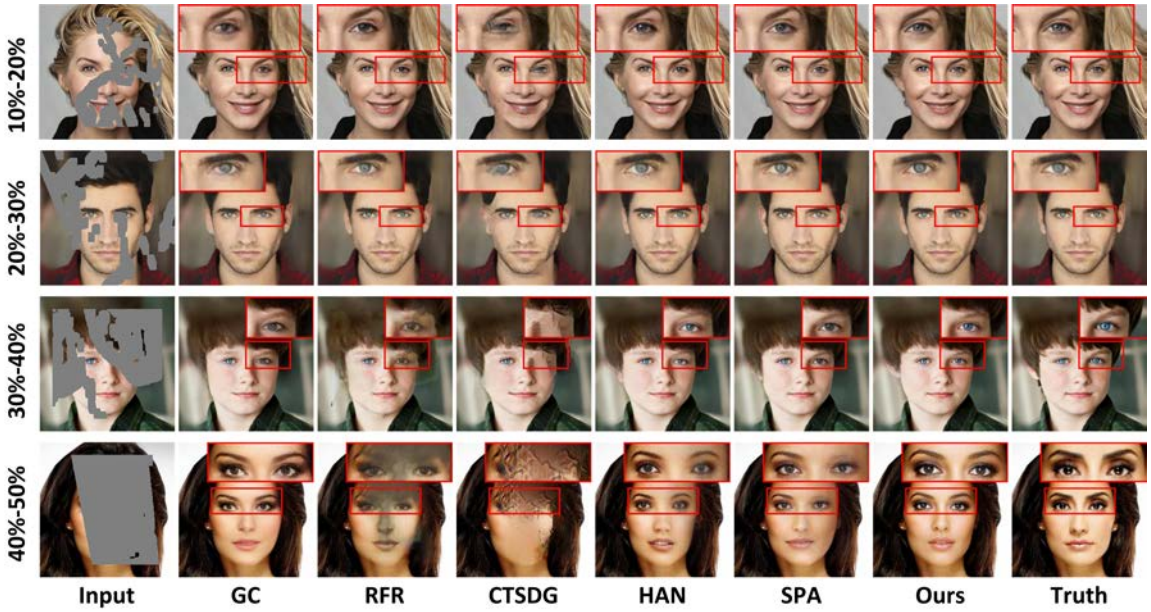


Fig. 6. Qualitative comparison with GC [10], RFR [7], CTSDG [38], HAN [39], and SPA [40] on the CelebA-HQ dataset. The first and second rows exhibit mask rates of 10%-20% and 20%-30% respectively, while the remaining rows are generated with mask rates of 30%-40% and 40%-50%. Zooming in on the red box shows a better view of the local details.

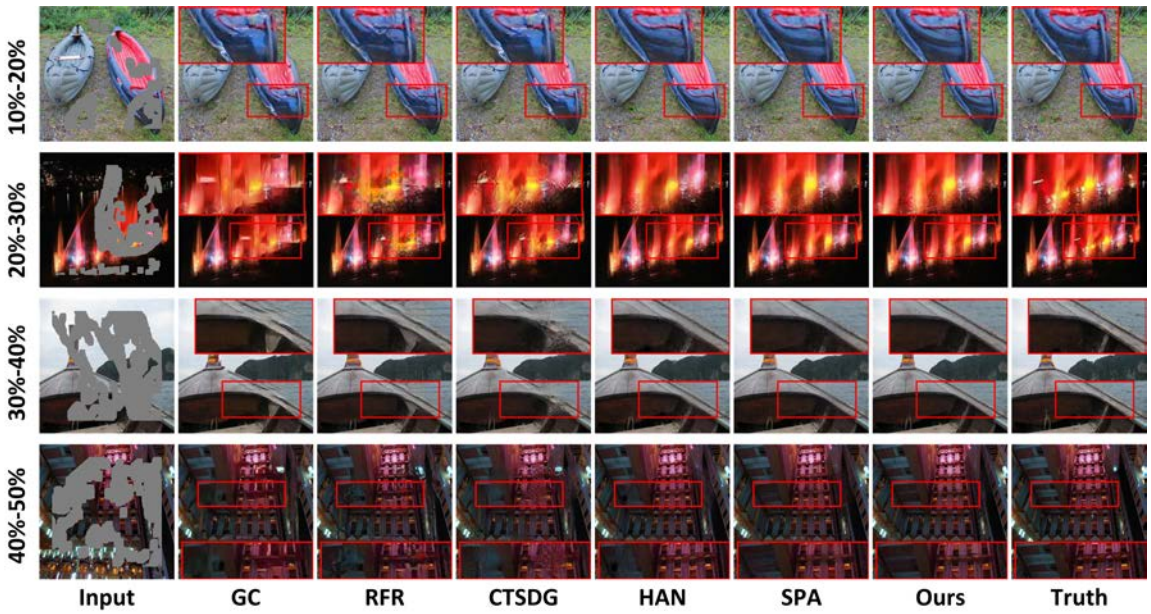


Fig. 7. Qualitative comparison with GC [10], RFR [7], CTSDG [38], HAN [39], and SPA [40] on the Places2 dataset. The first and second rows exhibit mask rates of 10%-20% and 20%-30% respectively, while the remaining rows are generated with mask rates of 30%-40% and 40%-50%. Zooming in on the red box shows a better view of the local details.

hierarchical feature fusion architecture, which enhances both efficiency and inpainting accuracy.

### *F. Ablation Study*

On the Paris StreetView, we analyze the effectiveness of each of the proposed modules, as shown in Table V. The term w/o DSGF denotes the removal of the dual-stream gated feature fusion module. Similarly, w/o RSAM refers to the exclusion of the region-separated attention module, and w/o MHLA indicates the replacement of efficient multi-head linear attention with traditional multi-head attention. The baseline model consists of the transformer encoder-decoder architecture, with both the DSGF and RSAM removed, and the MHLA replaced by traditional multi-head attention. The experimental results highlight the critical role of each component, as their removal significantly degrades inpainting performance, underscoring their importance to the model's effectiveness. The MHLA module has the most significant impact, as it more effectively captures long-range dependencies through explicit interactions

TABLE IV
COMPUTATIONAL COMPLEXITY COMPARISON RESULTS IN TERMS OF PARAMETERS (PARAMS) AND FLOATING-POINT OPERATIONS (FLOPS).

| Methods | Params | FLOPs |
|---|---|---|
| GC [10] | 16.0 M | 103.1 G |
| RFR [7] | 31.2 M | 172.1 G |
| CTSDG [38] | 52.1 M | 126.7 G |
| HAN [39] | 19.4 M | 137.7 G |
| SPA [40] | 13.2 M | 46.8 G |
| Ours | 16.1 M | 61.55 G |

with global features. The RSAM module is the second most effective, as it better captures the semantic information of the image, leading to more accurate inpainting. As expected, the DSGF module provides enhanced multi-level feature learning, enabling the learning of contextual feature representations that reduce artifacts and distorted edges. These results demonstrate that the DSGF, RSAM, and MHLA modules significantly improve performance, enabling the HFCT to capture useful interaction information and produce better inpainting results.

TABLE V
ABLATION STUDIES OF DIFFERENT MODULES ON THE PARIS STREETVIEW. ↓ MEANS LOWER IS BETTER, ↑ MEANS HIGHER IS BETTER.

| Mask Ratio | | 10%-20% | 20%-30% | 30%-40% | 40%-50% |
|---|---|---|---|---|---|
| LPIPS↓ | w/o DSGF | 0.023 | 0.043 | 0.066 | 0.097 |
| | w/o RSAM | 0.024 | 0.044 | 0.068 | 0.100 |
| | w/o MHLA | 0.024 | 0.044 | 0.068 | 0.100 |
| | Baseline | 0.028 | 0.050 | 0.076 | 0.113 |
| | Ours | **0.022** | **0.041** | **0.064** | **0.094** |
| PSNR↑ | w/o DSGF | 32.754 | 29.656 | 27.340 | 25.441 |
| | w/o RSAM | 32.713 | 29.729 | 27.425 | 25.438 |
| | w/o MHLA | 32.630 | 29.552 | 27.266 | 25.357 |
| | Baseline | 31.758 | 28.845 | 26.703 | 24.720 |
| | Ours | **33.022** | **29.961** | **27.622** | **25.693** |
| SSIM↑ | w/o DSGF | 0.952 | 0.913 | 0.865 | 0.806 |
| | w/o RSAM | 0.952 | 0.913 | 0.866 | 0.808 |
| | w/o MHLA | 0.951 | 0.912 | 0.864 | 0.803 |
| | Baseline | 0.945 | 0.903 | 0.854 | 0.790 |
| | Ours | **0.954** | **0.916** | **0.870** | **0.814** |
| FID↓ | w/o DSGF | 14.398 | 26.013 | 36.343 | 50.018 |
| | w/o RSAM | 14.522 | 25.340 | 36.885 | 48.062 |
| | w/o MHLA | 14.693 | 26.000 | 36.955 | 48.863 |
| | Baseline | 18.028 | 30.301 | 42.427 | 58.138 |
| | Ours | **13.552** | **23.598** | **34.644** | **47.388** |

## V. CONCLUSION

In this paper, we propose an efficient hierarchical feature collaboration transformer (HFCT) model for image inpainting, addressing insufficient detail reconstruction. HFCT employs a uniquely designed dual-stream gated feature fusion module (DSGF) and a region-separated attention module (RSAM) to effectively integrate multi-level features. DSGF enhances global consistency and reduces artifacts, while RSAM captures feature associations to refine results. This strategy enables the full utilization of multi-layer information, solving the layer interaction issues seen in existing models. HFCT performs well on the Paris StreetView, CelebA-HQ, and Places2 datasets, particularly at high masking rates and in FID metrics. It excels at processing complex images, maintaining coherent lines, and restoring detailed features like eyes, leading to realistic results. Limitations include its reliance on specific datasets and potential computational inefficiency, highlighting the need for future work on generalization and efficiency improvements.

## REFERENCES


[1] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, 2000, pp. 417–424.

[2] T. F. Chan and J. Shen, "Nontexture inpainting by curvature-driven diffusions," *Elsevier Journal of Visual Communication and Image Representation*, vol. 12, no. 4, pp. 436–449, 2001.

[3] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, "Filling-in by joint interpolation of vector fields and gray levels," *IEEE Transactions on Image Processing*, vol. 10, no. 8, pp. 1200–1211, 2001.

[4] H. Yamauchi, J. Haber, and H.-P. Seidel, "Image restoration using multiresolution texture synthesis and image inpainting," in *Proceedings Computer Graphics International 2003*, 2003, pp. 120–125.

[5] N. Komodakis and G. Tziritas, "Image completion using efficient belief propagation via priority scheduling and dynamic pruning," *IEEE Transactions on Image Processing*, vol. 16, pp. 2649–2661, 11 2007.

[6] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics*, vol. 36, pp. 1–14, 07 2017.

[7] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, "Recurrent feature reasoning for image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7760–7768.

[8] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 85–100.

[9] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2536–2544.

[10] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4471–4480.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[12] Z. Wan, J. Zhang, D. Chen, and J. Liao, "High-fidelity pluralistic image completion with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4692–4701.

[13] Y. Yu, F. Zhan, R. Wu, J. Pan, K. Cui, S. Lu, F. Ma, X. Xie, and C. Miao, "Diverse image inpainting with bidirectional and autoregressive transformers," *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 69–78, 2021.

[14] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, "Mat: Mask-aware transformer for large hole image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 758–10 768.

[15] M. Bertalmio, "Strong-continuation, contrast-invariant inpainting with a third-order optimal pde," *IEEE Transactions on Image Processing*, vol. 15, pp. 1934–1938, 07 2006.

[16] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Transactions on Image Processing*, vol. 13, pp. 1200–1212, 09 2004.

[17] J. Hays and A. A. Efros, "Scene completion using millions of photographs," *Communications of the ACM*, vol. 51, pp. 87–94, 10 2008.

[18] S. Darabi, E. Shechtman, C. Barnes, D. B. Goldman, and P. Sen, "Image melding," *ACM Transactions on Graphics*, vol. 31, pp. 1–10, 08 2012.

[19] J. Liu, S. Yang, Y. Fang, and Z. Guo, "Structure-guided image inpainting using homography transformation," *IEEE Transactions on Multimedia*, vol. 20, no. 12, pp. 3252–3265, 2018.

[20] J. Peng, D. Liu, S. Xu, and H. Li, "Generating diverse structure for image inpainting with hierarchical vq-vae," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 775–10 784.

[21] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2*, 2014, pp. 2672–2680.

[22] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi, "Edgeconnect: Generative image inpainting with adversarial edge learning," *ArXiv*, vol. abs/1901.00212, 2019.
[23] W. Xiong, J. Yu, Z. Lin, J. Yang, X. Lu, C. Barnes, and J. Luo, "Foreground-aware image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5840–5848.
[24] Q. Wang, Z. Wang, X. Zhang, and G. Feng, "Art image inpainting with style-guided dual-branch inpainting network," *IEEE Transactions on Multimedia*, vol. 26, pp. 8026–8037, 2024.
[25] J. Yang, Z. Qi, and Y. Shi, "Learning to incorporate structure knowledge for image inpainting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 12 605–12 612.
[26] H. Liu, B. Jiang, Y. Song, W. Huang, and C. Yang, "Rethinking image inpainting via a mutual encoder-decoder with feature equalizations," in *Proceedings of the European Conference on Computer Vision*. Springer, 2020, pp. 725–741.
[27] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.
[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[29] C. Zheng, T.-J. Cham, J. Cai, and D. Phung, "Bridging global context interactions for high-fidelity image completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 512–11 522.
[30] C. Zheng, G. Song, T.-J. Cham, J. Cai, D. Phung, and L. Luo, "High-quality pluralistic image completion via code shared vqgan," *arXiv preprint arXiv:2204.01931*, 2022.
[31] C. Cao, Q. Dong, and Y. Fu, "Learning prior feature and attention enhanced image inpainting," in *European Conference on Computer Vision*, 2022, pp. 306–322.
[32] S. Chen, A. Atapour-Abarghouei, and H. P. H. Shum, "Hint: High-quality inpainting transformer with mask-aware encoding and enhanced attention," *IEEE Transactions on Multimedia*, vol. 26, pp. 7649–7660, 2024.
[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.
[34] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 694–711.
[35] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2414–2423.
[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[37] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.
[38] X. Guo, H. Yang, and D. Huang, "Image inpainting via conditional texture and structure dual generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 134–14 143.
[39] Y. Deng, S. Hui, R. Meng, S. Zhou, and J. Wang, "Hourglass attention network for image inpainting," in *Proceedings of the European Conference on Computer Vision*, Cham, 2022, pp. 483–501.
[40] W. Huang, Y. Deng, S. Hui, Y. Wu, S. Zhou, and J. Wang, "Sparse self-attention transformer for image inpainting," *Pattern Recognition*, vol. 145, pp. 109 897–109 897, 01 2024.
[41] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.
[42] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.
[43] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
[44] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

**Dengyong ZHANG** received the B.S. and M.S. degree from Changsha University of Science and Technology, Changsha, China, in 2003, 2006 respectively. He received Ph.D. degree from Hunan University, Changsha, China, in 2018. Now, He is a Professor at Changsha University of Science and Technology. His current research interests include digital media forensics and image processing. (Email: zhdy@csust.edu.cn)

**Nuo Fu** received the B.E. degree from Southwest University of Science and Technology, Mianyang, China, in 2022. She is currently pursuing the M.S. degree at School of Computer and Communication Engineering, Changsha University of Science and Technology, in Hunan. Her research interests include image inpainting and image processing. (Email: 22208051655@stu.csust.edu.cn)

**Xin Liao** eceived the B.E. and Ph.D. degrees in information security from the Beijing University of Posts and Telecommunications in 2007 and 2012, respectively. He is currently a Professor and a Doctoral Supervisor with Hunan University, China. He worked as a Post-Doctoral Fellow with the Institute of Software, Chinese Academy of Sciences, and also a Research Associate with The University of Hong Kong. From 2016 to 2017, he was a Visiting Scholar with the University of Maryland, College Park, USA. His current research interests include multimedia forensics, steganography, and watermarking. (Email: xinliao@hnu.edu.cn)

**Jiaxin CHEN** received the B.S. degree from Central China Normal University, Wuhan, China, in 2017, and Ph.D. degree from Hunan University, Changsha, China in 2023. She is currently a Lecturer with Changsha University of Science and Technology, China. Her current research interests include multimedia forensics and Deepfake detection. She is a member of Technical Committee (TC) on Digital Forensics and Security of China Society of Image and Graphics. (Email: jxchen@csust.edu.cn)

**Hengfu Yang** received the M.S. degree from Guizhou University, in 2003, and the PhD degree from Hunan University, China, in 2009. He is now a professor at the School of Computer Science, Hunan First Normal University, China. His main research interests include information hiding, image clustering, digital forensics, privacy-preserving, and multimedia encryption. (Email: hengfuyang@hnfnu.edu.cn)

**Gaobo Yang** received the B.S. and M.S. degree in 1995 and 2001 respectively from Shenyang University of Technology and East China Jiaotong University, China. He received the Ph.D. from Shanghai University, china, in 2004. Currently he is a professor in Hunan University, China. his research interests are in the area of image and video signal processing, digital media forensics. (Email: yanggaobo@hnu.edu.cn)