SyFormer: Structure-Guided Synergism Transformer for Large-Portion Image Inpainting

Jie Wu[1], Yuchao Feng[1], Honghui Xu[1], Chuanmeng Zhu[2], Jianwei Zheng[1]*,

[1]Zhejiang University of Technology
[2]Zhejiang University
wuj@zjut.edu.cn, fyc@zjut.edu.cn, xhh@zjut.edu.cn, cmzhu@zju.edu.cn, zjw@zjut.edu.cn

Abstract

Image inpainting is in full bloom accompanied by the progress of convolutional neural networks (CNNs) and transformers, revolutionizing the practical management of abnormity disposal, image editing, etc. However, due to the ever-mounting image resolutions and missing areas, the challenges of distorted long-range dependencies from cluttered background distributions and reduced reference information in image domain inevitably rise, which further cause severe performance degradation. To address the challenges, we propose a novel large-portion image inpainting approach, namely the Structure-Guided Synergism Transformer (SyFormer), to rectify the discrepancies in feature representation and enrich the structural cues from limited reference. Specifically, we devise a dual-routing filtering module that employs a progressive filtering strategy to eliminate invalid noise interference and establish global-level texture correlations. Simultaneously, the structurally compact perception module maps an affinity matrix within the introduced structural priors from a structure-aware generator, assisting in matching and filling the corresponding patches of large-proportionally damaged images. Moreover, we carefully assemble the aforementioned modules to achieve feature complementarity. Finally, a feature decoding alignment scheme is introduced in the decoding process, which meticulously achieves texture amalgamation across hierarchical features. Extensive experiments are conducted on two publicly available datasets, i.e., CelebA-HQ and Places2, to qualitatively and quantitatively demonstrate the superiority of our model over state-of-the-arts.

Introduction

Given corrupted images, inpainting technique endeavors to restore obscure regions with semantically coherent contents. Propelled by rapid advancements in digital media, this technique has garnered significant attention and found wide-ranging applications in real-world scenarios, such as picture editing (Song et al. 2019), restoration of polluted images(Jin et al. 2023; Feihong et al. 2023), and disposal of undesired objects (Li, Wang, and Hu 2021). However, with ongoing iterations of hardware and software, the captured images enjoy a consistently higher resolution, which poses new challenges to the mission. In practice, high resolution implies more information as well as more complex properties, whereas the input incompleteness inevitably suffers drastic changes in feature distribution, which further leads to cluttered image domain and brings about a catastrophic performance drop for existing methods. Especially in cases with large-proportion pixel loss, most previous research loses its original luster due to the severe scarcity of available references.

(a) (b) (c)

Figure 1: Inpainted results in cases of background disturbance (upper row) and large missing regions (bottom row). (a) Corrupted images. (b) Inpainted results. (c) GT.

To address these challenges, massive studies based on convolutional neural networks (CNNs) (Xiong et al. 2019; Jin et al. 2023) have come up with well-thought-out ideas. The main solution is to add auxiliary priors to guide pixel filling on top of the powerful learning capability of CNNs, which proves pivotal in mitigating the issue of limited information within the occluded regions. In particular, the auxiliary information is typically inspired by contextual cues, e.g., edge maps (Nazeri et al. 2019), semantic maps (Liao et al. 2020), etc. Following this line, most current approaches bifurcate the inpainting process into two branches. The initial one revolves around the latent cues associated with the impaired regions, succeded by the intricate texture synthesis. For example, CTSDG (Guo, Yang, and Huang 2021) develops a parallel architecture that simultaneously

*Corresponding author

models intermediate prior generation and texture inpainting. Additionally, Lama (Suvorov et al. 2022) introduces an innovative Fast Fourier Convolution (FFC) module, harnessing frequency domain transformations to extract high receptive field structures at minimal computational expense. Despite the integration of auxiliary priors to mitigate the challenge of data scarcity, these methods remain insufficient to fully excavate the precise information concealed within intricate image distributions. Furthermore, they are susceptible to entrapment within localized matching predicaments. This phenomenon is attributed to the intrinsic inductive bias inherent in CNNs, which restricts their efficacy in capturing long-range dependencies. In light of these insights, there exists significant scope unexplored for the more comprehensive apprehension of global image distributions.

In contrast to CNNs, vision transformer (Dosovitskiy et al. 2020) has emerged as a formidable instrument for modeling non-local dependencies, sparking a revolution in the community of computer vision. Leveraging the power of self-attention, researchers have devised tailored approaches (Yu et al. 2021; Wan et al. 2021) to achieve high-fidelity inpainting outcomes. However, these methods show appealing performance only when dealing with low-resolution images, e.g., 256×256. High-quality inpainting of incoherent background textures remains a challenge when faced with high-resolution inputs, e.g., 512×512. For instance, MAT (Li et al. 2022) exclusively uses pertinent tokens to model the contextual information during the low resolution stage, followed by the dynamic update stage. Yet, it still leaves the question of whether low-resolution self-attention can be perfectly matched to high-resolution scenes. On the contrary, ZITS (Dong, Cao, and Fu 2022) uses transformer directly to obtain holistic edges rather than acting on the detailed inpainting process. Although these treatments have opened a new avenue for high-resolution inpainting, the efficient modeling of the global incoherent textures, especially in cases with large-proportion occlusions, is still in short consideration. Moreover, transformer-based approaches also face resource constraints when the input resolution increases, as the attention mechanism necessitates pairwise affinity computations across all spatial positions, resulting in exorbitant computational complexity and memory requirements. This pressing concern necessitates the formulation of a judicious methodology capable of effectively tackling high-resolution image-inpainting scenarios. Some studies (Zhu et al. 2023) focus on improving the functioning of transformers to tackle the problems mentioned above. In their self-attention calculations, each query is linked to a restricted number of key-value pairs, thus decreasing the computational load.

Inspired by these insights, we propose a new large-portion inpainting approach, namely Structure-Guided Synergism Transformer (SyFormer), which taps the potential of modeling comprehensive contexts and capturing prototypical cues. To achieve this, we decouple image inpainting into two stages: 1) A structure-aware generator is deployed, which together with a Contextual Attention Block (CAB) engenders coarse representations of the impaired inputs. In contrast to the predominant techniques that rely on edge maps as reconstruction priors (Nazeri et al. 2019), our approach crafts structure maps imbued with richer semantics, thus mitigating the intrinsic information paucity endemic to image domains. 2) A synergism transformer is proposed to simultaneously process the corrupted image and the structure map in a parallel symphony, effectively resolving entanglements and enhancing the completion performance. More importantly, within the SyFormer framework, a Structurally Compact Perception (SCP) module plays the central role, by a lightweight self-attention mechanism, which molds the global correlations of initially generated features and then guides the adaptive matching of corresponding patches from the original image. In addition, a Dual-Routing Filtering (DRF) module is designed to dynamically appraise the region-wise affinity, which further aggregates a predetermined tally of region tokens and enables global modeling of complex image distributions. Note that the overall mechanism facilitates seamless collaboration between the two attention modules, affording an extensive exploration of the inherent pixel entwinements. Besides, for better texture quality, we delve into a Feature Decoding Alignment (FDA) methodology that effectively amalgamates feature information from hierarchical levels. As shown in Fig. 1, our SyFormer is capable of removing unwanted objects and filling in the large holes with visually authentic content. The main contributions are given as follows.

- We propose a dual-stage architecture to tackle the challenge of limited reference information due to large areas of missing pixels. The initial stage is dedicated to the extraction of structural cues as prior information. Then, a parallel-disentanglement framework is introduced, which concurrently processes the harmonious structure and corrupted images, enabling a meticulous examination and discernment of visual properties.
- To overcome the challenges posed by chaotic background distributions, a lightweight synergism transformer is proposed, whose main architecture can be divided into two parallel branches. In terms of the corrupted image, the lower branch introduces a dual-routing filtering module that employs a coarse-to-fine scheme to achieve specificity of large-proportion incoherent textures. In terms of structure map, the upper branch leverages a structurally compact perception module to capture contextual semantic information through global modeling, thereby guiding the matching of corresponding patches.
- To facilitate precise texture synthesis, a feature decoding alignment scheme is designed, which gradually blends the multiscale spatial-channel features to achieve the intact finals. Extensive experiments on CelebA-HQ (Karras et al. 2018) and Places2 (Zhou et al. 2017) datasets are conducted for assessment, whose results demonstrate that our model significantly outperforms state-of-the-art approaches, both qualitatively and quantitatively.

Related Work

Auxiliary-Prior Inpainting

Image inpainting, as a long-standing problem in computer vision, has made significant strides in recent years with the

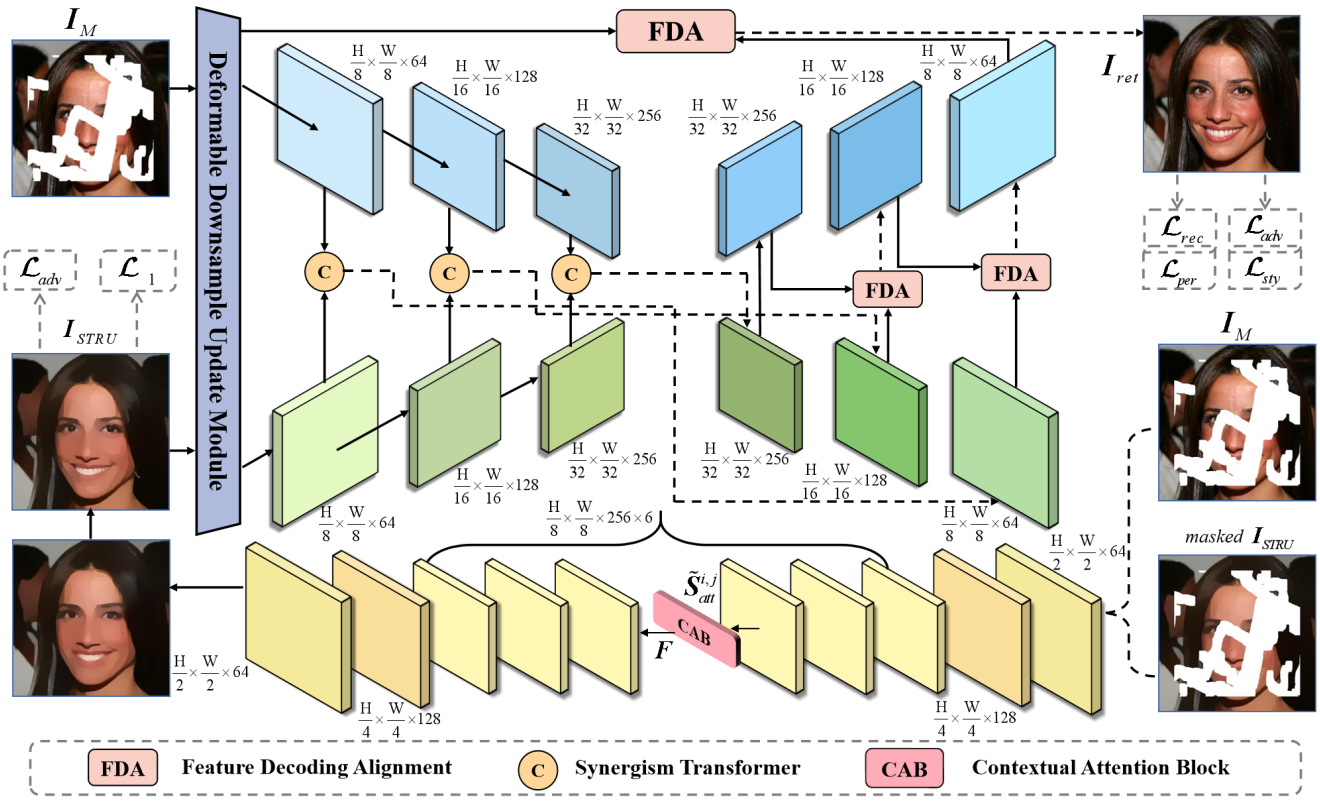


Figure 2: The overall architecture of SyFormer. A disentanglement architecture processes both the structure map and the original input in a parallel manner, allowing meticulous semantic examination of contextual features.

advent of deep learning techniques based on convolutional neural networks (CNNs) (Zuo et al. 2023; Pathak et al. 2016; Jin et al. 2022). In contrast to earlier approaches those depend heavily on hand-crafted priors (Xu et al. 2021, 2022), deep learning-based methods have demonstrated great potential in extracting hierarchical representations from images. This enables the reconstruction of damaged content with a high-level understanding of the underlying semantics. However, while remarkable performance boosts have been achieved, challenges still exist due to the inherent uncertainty when large proportions of pixels are missed. In addition, the absence of a well-defined constraint to facilitate a better convergence is also a daunting issue. To that end, various auxiliary sources of prior information have been explored, such as edge maps, foreground contours (Xiong et al. 2019), and structural guidance (Ren et al. 2019), which are currently known as the key ingredient for performance guarantee. Specifically, EdgeConnect (Nazeri et al. 2019) devised an edge generator to obtain contour sketches, which together with a texture generator contributes to a stable reconstruction. Analogously, CTSDG (Guo, Yang, and Huang 2021) introduced a dual-stream network to couple structure-constrained texture synthesis and texture-guided structure reconstruction. Using the power of FFC, Lama (Suvorov et al. 2022) achieves remarkable efficacy in high-resolution image inpainting tasks, successfully reconciling complex details and significant damage. In this work, we also employ the structure map as an auxiliary prior. However, rather than merely joining the auxiliary prior with the damaged image through concatenation, we map the structural prior to an affinity matrix to help fill in the damaged areas.

Visual Transformer Inpainting

Transformer has made remarkable achievements in the field of natural language processing (NLP) by virtue of the impressive capacity to model long-range dependencies. Recently, its applicability has been greatly extended to the realm of computer vision. For instance, Vision Transformer (ViT) (Dosovitskiy et al. 2020) demonstrated that pure transformer networks can achieve the same level of classification performance as the CNN-based counterparts. More recently, massive efforts have been dedicated to mitigate the heavy computational burden of self-attention, including discrete representation learning (Ramesh et al. 2021) and some linear treatments (Feng et al. 2023). Moreover, transformers have gained widespread adoption in the field of image inpainting, enjoying superior reconstruction quality with finer details. Representatively, Yu et al. (Yu et al. 2021) employed reordering techniques and autoregressive modeling to effectively capture global priors from both valid and masked pixels. Specific to the high-resolution challenges, MAT (Li et al. 2022) introduced a dynamic mask updating mechanism to incorporate long-range dependencies across relevant tokens. Rather than using a transformer to generate intermedi-

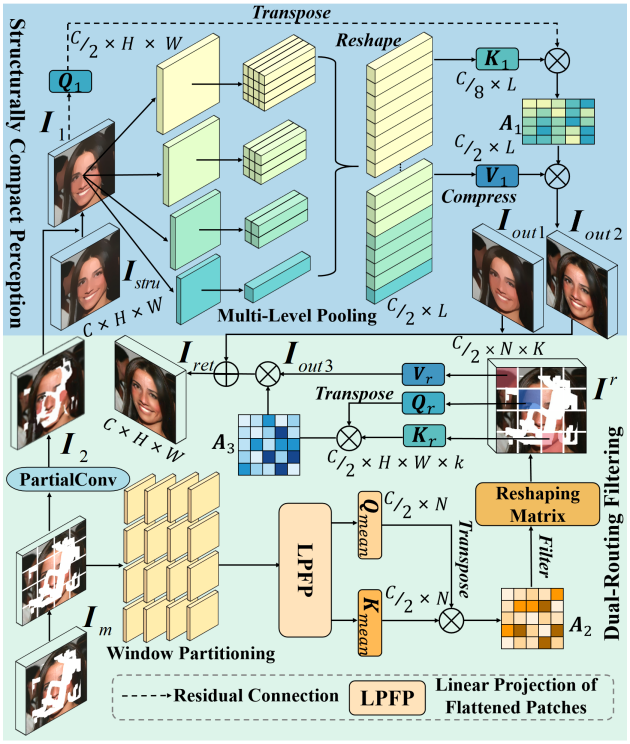


Figure 3: Illustration of the synergism transformer.

ate clues and then employing a CNN for texture filling, our research instead uses a transformer variant throughout the entire filling process. This enables us to create a parallelized disentanglement architecture that enriches information representations.

Methodology

The goal of inpainting is to restore the corrupted regions of a masked image, $\boldsymbol{I}_M = \boldsymbol{I} \odot \boldsymbol{M}$, by filling in visually plausible content that is consistent with the context, where $\boldsymbol{I}$ is the ground truth and $\boldsymbol{M}$ denotes the mask. As shown in Fig. 2, the Structure-Guided Synergism Transformer is made up of two parts: a structure-aware generator that produces rough sketches and a synergism transformer that generates realistic texture details. The U-shaped architecture of the structure-aware generator allows it to investigate the intrinsic connections within the image, enabling the production of structural maps $\boldsymbol{I}_{STRU}$. Taking into account both the coarse structure $\boldsymbol{I}_{STRU}$ and the corrupted image $\boldsymbol{I}_M$, the synergism transformer facilitates the synthesis of high-resolution fine-grained textures to produce the inpainted result $\boldsymbol{I}_{ret}$. In particular, our synergism transformer architecture further consists of two concurrents, i.e. structurally compact perception and dual-routing filtering. Finally, a feature decoding alignment technique is applied to amalgamate texture information across neighboring layers and rectify semantic discrepancies.

Structure-Aware Generator

The bottom segment of Fig. 2 provides the entire pipeline of structure-aware generator. The paired inputs of the masked image and an initial sketch are fed into a U-shaped encoder-decoder architecture. To discover the latent possibilities of coherent structure within damaged regions, encoded low resolution features undergo the contextual attention block (CAB) (Yu et al. 2018), which facilitates the learning of an affinity matrix surrounding damaged regions, thus enhancing the extraction of attention for potential texture patches.

Following the mechanism of CAB, we first extract 3×3 patches both from the intact and corrupted region of the encoded feature map $\boldsymbol{F}$, and then compute the cosine similarity between the patches:

$$\boldsymbol{S}_{att}^{i,j} = \left\langle \frac{\boldsymbol{p}_i}{\|\boldsymbol{p}_i\|_2}, \frac{\boldsymbol{p}_j}{\|\boldsymbol{p}_j\|_2} \right\rangle, \tag{1}$$

where $\boldsymbol{p}_i$ is the i-th patch taken from the intact region, while $\boldsymbol{p}_j$ is the j-th patch of the corrupted region. The attention score of each patch is then calculated by applying softmax to the cosine similarity $\boldsymbol{S}_{att}^{i,j}$:

$$\tilde{\boldsymbol{S}}_{att}^{i,j} = \frac{\exp(\boldsymbol{S}_{att}^{i,j})}{\sum_{j=1}^{N} \exp(\boldsymbol{S}_{att}^{i,j})}, \tag{2}$$

where N is the number of patches outside the missing hole.

After obtaining the attention scores from $\boldsymbol{F}$, the deconvolution operation is used to fill the corresponding holes of the low-level feature maps with weighted contextual patches. The process can be formulated as follows:

$$\hat{\boldsymbol{p}}_i = \sum_{j=1}^{N} \hat{\boldsymbol{p}}_j \tilde{\boldsymbol{S}}_{att}^{i,j}, \tag{3}$$

where $\hat{\boldsymbol{p}}_j$ denotes a 3×3 patch extracted from low-level feature maps $\boldsymbol{F}_l$, and $\hat{\boldsymbol{p}}_i$ is the newly learned patch.

Synergism Transformer

Structurally Compact Perception Note again that the primary aim of image inpainting is to systematically search for prospective pixels that fill the masked regions. However, conventional deep inpainting methods tend to encounter obstacles such as insufficient in-domain information for accurate pixel recovery in cases of large-portion pixel missing. Accordingly, to effectively guide pixel matching within the damaged regions and mitigate interference from mask-related noise, we incorporate the structure map $\boldsymbol{I}_{STRU}$ generated in the first stage as an extra guide. In practice, as illustrated in Fig. 3, we conceive the SCP module that is rooted in typical self-attention mechanism (as given in Eq. (4)):

$$selfAttention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Softmax(\boldsymbol{Q} \cdot \boldsymbol{K}^T) \cdot \boldsymbol{V}, \tag{4}$$

where sequences $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in R^{N \times C}$ are attained by three learnable weight matrices, $N = H \times W (C \ll N)$, H and W are the height and width of the features, C is the channel dimension, and T represents the transpose operation. It is notorious that Eq. (4) suffers from $O(N^2)$ space and time

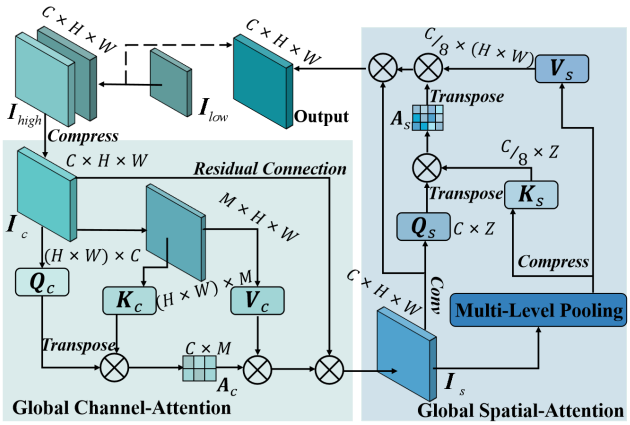


Figure 4: Feature decoding alignment block.

complexity, which imposes substantial resource consumption, particularly in cases of high-resolution inputs.

To control the computational consumption, we commence by employing the Deformable Downsample Update Module (Jin et al. 2023) to extract multi-scale feature information, $\boldsymbol{I}_m$ and $\boldsymbol{I}_{stru}$, from the original inputs, $\boldsymbol{I}_M$ and $\boldsymbol{I}_{STRU}$, respectively. For the structural image $\boldsymbol{I}_{stru}$, a $C/2$ convolutional kernel with reduced channel dimension is used to access the intrinsic feature $\boldsymbol{I}_1$. Furthermore, $\boldsymbol{I}_1$ is augmented with an additional map of auxiliary features $\boldsymbol{I}_2$, earning for the progressive restoration through partial convolution operations of $\boldsymbol{I}_m$. The introduction of the progressively enriched features $\boldsymbol{I}_2$ plays the role as adding an adversarial convolutional process between newly filled pixels and the existing structural pixels, thus improving the effectiveness of the prior cues. Subsequently, to achieve enhanced feature compression, we incorporate the operations of multi-level pooling and linear projections. At this juncture, three feature sequences $\boldsymbol{K}_1 \in R^{C/8\times L}$, $\boldsymbol{V}_1 \in R^{C/2\times L}$ and $\boldsymbol{Q}_1 \in R^{C/2\times(H\times W)}$ $(L \ll H \times W)$ are acquired that have undergone compression. Then, specific self-attention scores $\boldsymbol{A}_1$ can be obtained using $\boldsymbol{Q}_1$ and $\boldsymbol{K}_1$:

$$\boldsymbol{A}_1 = \boldsymbol{Q}_1^T \cdot \boldsymbol{K}_1 = (Conv(\boldsymbol{I}_1))^T \cdot Com\left(\varepsilon\left(\boldsymbol{I}_1\right)\right), \quad (5)$$

where $Conv\left(\cdot\right)$ and $Com\left(\cdot\right)$ stand for convolution and compression operations, respectively. $\varepsilon\left(\cdot\right)$ denotes the multi-level pooling. Finally, the feature map $\boldsymbol{I}_{out1}$ is derived by aggregating global structural information and semantic discriminative features:

$$\boldsymbol{I}_{out1} = SCP\left(\boldsymbol{I}_1\right) = \varepsilon\left(\boldsymbol{I}_1\right) \cdot Softmax(\boldsymbol{A}_1)^T + \boldsymbol{I}_1. \quad (6)$$

Likewise, the progressively inpainted feature map $\boldsymbol{I}_{out2}$ can be acquired through the same procedure from $\boldsymbol{I}_2$.

Dual-Routing Filtering Current methods often directly apply transformer to incomplete inputs, overlooking the fact that the fundamental nature of transformer is to discover interdependencies between tokens. Consequently, patches with a substantial number of missing pixels would cause pronounced errors during self-attention computations. To establish the inter-patch connections and meanwhile avoid the interference caused by mask noise, we design a dual-routing filtering module, as depicted in the lower segment of Fig. 3. Initially, to control the computational burden, the channel dimension of the corrupted input is reduced through a convolution operation. Afterwards, operations of window partition and weight mapping are executed, resulting in the generation of non-overlapping sets of query, key, and value tensors, $\boldsymbol{Q}^w, \boldsymbol{K}^w, \boldsymbol{V}^w \in R^{HW/s^2\times C/2}$. These operations enable the model to focus on valuable tokens over long distances, thereby fostering robust contextual relationships. The tensors $\boldsymbol{Q}^w$ and $\boldsymbol{K}^w$ are then put through linear projection of flattened patches (LPFP) and average operations to obtain the region-level responses $\boldsymbol{Q}_{mean}$, $\boldsymbol{K}_{mean}$, which reduce the interference from large amount of masking. Finally, the adjacency matrix $\boldsymbol{A}_2$, representing the region-level affinity graph, can be calculated using the following formula:

$$\boldsymbol{A}_2 = Softmax\left(\boldsymbol{K}_{mean} \cdot (\boldsymbol{Q}_{mean})^T\right). \quad (7)$$

Based on $\boldsymbol{A}_2$, most of the ineffective areas can be filtered out, with the more pertinent regions reserved. To that end, we transform $\boldsymbol{A}_2$ with a gating threshold to a sparse affinity matrix $\boldsymbol{F}^r$, allows us to reshape a more refined ourcome of the original image. The process is described as follows:

$$\boldsymbol{I}^r = \delta\left(\boldsymbol{F}^r\right) = \delta\left(Filter\left(\boldsymbol{A}_2\right)\right), \quad (8)$$

where $\delta\left(\cdot\right)$ denotes the reshaping operation. On that basis, each query establishes strong ties with tokens from the most pertinent areas. By exclusively considering elements within these regions and performing aggregation operations, DRF is able to achieve global pixel matching, particularly in scenarios with inconsistent textures. Simultaneously, the implementation of this filtering-aggregation mechanism successfully mitigates the computational burden of transformer. Specifically, we perform aggregation of pixels from affinity regions to derive the corresponding $\boldsymbol{K}_r \in R^{s^2\times kHW/s^2\times C/2}$ and $\boldsymbol{V}_r \in R^{s^2\times kHW/s^2\times C/2}$. The resultant feature map $\boldsymbol{I}_{out3}$ can be derived by computing the following self-attention formula:

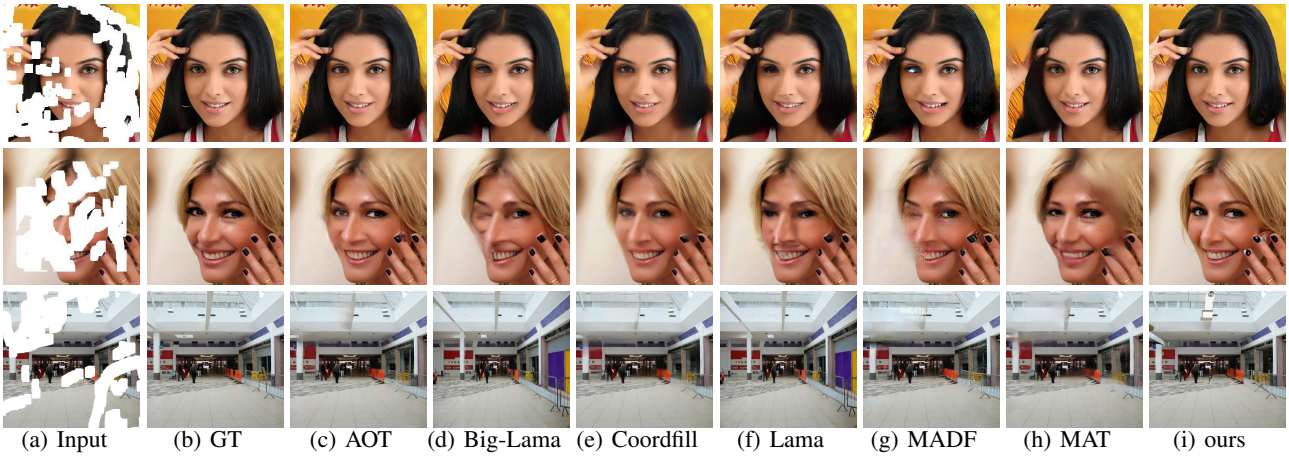
$$\boldsymbol{I}_{out3} = \boldsymbol{V}_r \cdot \boldsymbol{A}_3^T = \boldsymbol{V}_r \cdot Softmax\big(\boldsymbol{Q}_r^T \cdot \boldsymbol{K}_r\big)^T. \quad (9)$$

Ultimately, the overall outcome can be generated by

$$\boldsymbol{I}_{ret} = \sigma\left(Conv\left(Concat\left(\boldsymbol{I}_{out1}, \boldsymbol{I}_{out2}, \boldsymbol{I}_{out3}\right)\right)\right), \quad (10)$$

where $\sigma\left(\cdot\right)$ denotes the Tanh activation function.

Feature Decoding Alignment The intricate details present in the multi-level features from SyFormer, denoted as $\boldsymbol{I}_{ret}$, are crucial for the successful decoding of realistic images. Among them, lower-level features with high resolution $\boldsymbol{I}_{high}$ offer the more refined textures, but may contain inconsistent content that leads to inaccurate sampling results. On the other hand, higher-level features $\boldsymbol{I}_{low}$ encompass a wide field of view and contain a lot of semantic information, which also assists in refining textures. That is to say, purely using low-level features or simply



(a) Input (b) GT (c) AOT (d) Big-Lama (e) Coordfill (f) Lama (g) MADF (h) MAT (i) ours

Figure 5: Qualitative comparisons on representative examples from CelebA-HQ and Places2. (a) Corrupted inputs, (b) GT images, (c) AOT, (d) Big-Lama, (e) Coordfill, (f) Lama, (g) MADF, (h) MAT, and (i) our SyFormer.

upsampling high-level features both cause severe distortions due to the limited information during the decoding phase.

Following these perspectives, efficiently matching multi-level features become a pivotal step in promoting the inpaiting quality. Thus, we introduce feature decoding alignment technique as shown in Fig. 4. Specifically, the size of $\boldsymbol{I}_{low}$ is firstly upsampled to the resolution of $\boldsymbol{I}_{high}$. Afterwards, to generate lightweight sequences, $\boldsymbol{K}_c, \boldsymbol{V}_c \in R^{M\times(H\times W)}$, and $\boldsymbol{Q}_c \in R^{C\times(H\times W)}$, a series of convolutions and linear projection operations are used. Following this, Eq. (4) is used to calculate channel-level attention scores $\boldsymbol{A}_c$, which is further used to generate feature $\boldsymbol{I}_s$. Moreover, to establish multi-scale channel-spatial relationships, global spatial-attention operations are applied to $\boldsymbol{I}_s$. Specifically, $\boldsymbol{K}_s$, $\boldsymbol{Q}_s$, and $\boldsymbol{V}_s$ are generated using analogous feature mapping and linear transformation. Drawing insights from the self-attention computation in Eq. (4), a feature response output characterized by both spatial and channel-level attributes is achieved.

Loss Functions For rendering visual authenticity and global continuity, our loss function is a mixture of perceptual loss, style loss, reconstruction loss, and adversarial loss.

- **Perceptual Loss** Perceptual Loss is implemented by the pre-trained VGG16 model, which can optimize the global structure of the image. The specific formula with l_1 norm is:

$$\boldsymbol{\mathcal{L}_{per}} = \sum_r \left\|\phi_r\left(\boldsymbol{I}_{ret}\right) - \phi_r\left(\boldsymbol{I}\right)\right\|_1, \tag{11}$$

  where $\phi_r()$ denotes the r-th pooling layer of VGG16, and the value range of pool r is $\{1, 2, 3\}$ in our implementation.

- **Style Loss** Style consistency is controlled with the following constraint with $\varphi_r\left(\cdot\right) = \phi_r^T\left(\cdot\right)\phi_r\left(\cdot\right)$.

$$\boldsymbol{\mathcal{L}_{sty}} = \sum_r \left\|\varphi_r\left(\boldsymbol{I}_{ret}\right) - \varphi_r\left(\boldsymbol{I}\right)\right\|_1. \tag{12}$$

- **Reconstruction Loss** The standard reconstruction loss is defined as the average absolute error of $\boldsymbol{I}_{ret}$ and the truth image $\boldsymbol{I}$, which easily results in problems such as artifacts, distorted contours, etc. Thus, we update the reconstruction loss as follows:

$$\boldsymbol{\mathcal{L}_{rec}} = \sum_{x,y\in M} \left\|\boldsymbol{I}_{ret}^{x,y} - \boldsymbol{I}^{x,y}\right\|_1 + \sum_{x,y\notin M} \left\|\boldsymbol{I}_{ret}^{x,y} - \boldsymbol{I}^{x,y}\right\|_1, \tag{13}$$

  where $x, y \in M$ denotes the location (x, y) in the corrupted region M.

- **Adversarial Loss** To generate the more realistic details, adversarial loss is also involved, such as:

$$\boldsymbol{\mathcal{L}_{adv}} = \mathrm{E}\left[\log\left(1 - \mathcal{D}_w\left(\boldsymbol{I}_{ret}\right)\right)\right] + \mathrm{E}\left[\log \mathcal{D}_w\left(\boldsymbol{I}\right)\right], \tag{14}$$

  where $\mathcal{D}$ is the discriminator parameterized by w.

  With all the sub-loss prepared, the final $\boldsymbol{\mathcal{L}_{hybrid}}$ can be expressed as:

$$\boldsymbol{\mathcal{L}_{hybrid}} = \lambda_{per}\boldsymbol{\mathcal{L}_{per}} + \lambda_{rec}\boldsymbol{\mathcal{L}_{rec}} + \lambda_{sty}\boldsymbol{\mathcal{L}_{sty}} + \lambda_{adv}\boldsymbol{\mathcal{L}_{adv}}, \tag{15}$$

  where λ_{per}, λ_{rec}, λ_{sty}, and λ_{adv} are all hyper-parameters for balancing the contribution of each sub-loss function.

Experiments

Datasets and Metrics

Two well-known datasets, i.e., CelebA-HQ (Karras et al. 2018) and Place2 (Zhou et al. 2017), are used for the performance investigation. Images of these datasets cover a wide range of scenes and contents, allowing us to train a more general model for real-world applications. The CelebA-HQ data is split into training, validation, and test sets in a ratio of 24:1:5. We keep 220,000 and 5000 images from the original places2 sets for training and testing, respectively. The spatial resolution of these images are adjusted to 512×512 by cropping and scaling. Similar to (Liu et al. 2018), 6000 irregular masks are generated, whose covering area occupies 30%-60% of the total image. In addition, the masks are equally

Metric		SSIM(↑)			PSNR(↑)			MAE(↓)			LPIPS(↓)		
Mask Ratio		30-40%	40-50%	50-60%	30-40%	40-50%	50-60%	30-40%	40-50%	50-60%	30-40%	40-50%	50-60%
CelebA-HQ	AOT	**0.9193**	0.8774	0.7792	27.4578	25.2925	22.0181	0.0193	0.0278	0.0471	0.0767	0.1089	0.1707
	Lama	0.8996	0.8508	0.7426	26.4725	24.4377	21.3557	0.0258	0.0351	0.0565	0.1009	0.1369	0.2017
	MAT	0.9098	0.8659	0.7698	26.8392	24.7488	21.6119	0.0204	0.0292	0.0482	0.0779	**0.1082**	0.1644
	MADF	0.8958	0.8437	0.7285	26.3758	24.2502	20.7712	0.0224	0.0325	0.0565	0.1137	0.1594	0.2374
	Coordfill	0.9151	0.8739	0.7781	27.5136	25.5431	22.4907	0.0196	0.0277	0.0454	0.0887	0.1217	0.1802
	Big-Lama	0.9022	0.8511	0.7221	26.6291	24.4942	20.9228	0.0246	0.0335	0.0583	0.0921	0.1292	0.2107
	ours	0.9184	**0.8818**	**0.8001**	**27.9467**	**26.0252**	**22.6334**	**0.0187**	**0.0259**	**0.0408**	**0.0763**	0.1084	**0.1640**
Places2	AOT	0.8291	0.7668	**0.6746**	23.3894	21.4761	18.8301	0.0328	0.0465	0.0726	0.1214	0.1691	0.2562
	Lama	0.8483	0.7841	0.6581	23.4259	21.7237	19.3171	0.0389	0.0506	0.0742	0.1245	0.1689	0.2534
	MAT	0.8156	0.7377	0.6001	21.9216	20.0850	17.4618	0.0382	0.0539	0.0839	0.1337	0.1841	0.2675
	MADF	0.8405	0.7705	0.6403	23.1857	21.4208	18.9999	0.0332	0.0463	0.0702	0.1384	0.1917	0.2823
	Coordfill	0.8498	0.7844	0.6634	23.7492	22.0542	19.7105	0.0315	0.0436	0.0663	0.1324	0.1801	0.2531
	Big-Lama	0.8529	0.7905	0.6663	23.6102	21.9125	19.5266	0.0379	0.0491	0.0717	0.1172	**0.1591**	**0.2397**
	ours	**0.8552**	**0.7922**	0.6695	**24.0885**	**22.3871**	**19.9352**	**0.0314**	**0.0431**	**0.0639**	**0.1161**	0.1685	0.2423

Table 1: Quantitative comparison on CelebA-HQ and Places2 datasets (↑Higher is better, ↓Lower is better). The best values are highlighted by boldface and the second-best values are highlighted by underlines.

divided into three parts according to different hole-to-image ratios, *e.g.*, 30-40%, 40-50%, and 50-60%.

Experimental Settings

All experiments are conducted using PyTorch with a batch size of 8. Our model is optimized by Adam with a learning rate of 2×10^{-4}. The hyper-parameters in Eq. (15) are set as $\lambda_{adv} = 0.1$, $\lambda_{rec} = 40$, $\lambda_{sty} = 120$, $\lambda_{per} = 0.05$ to generate the sensuously optimal results. All experiments are conducted on two GPUs of RTX 3090 with a single 12G of video memory. To substantiate the efficacy of our proposal, we conduct a comprehensive comparison against state-of-the-art inpainting methods. Specifically, AOT (Zeng et al. 2023), MAT (Li et al. 2022), Coordfill (Liu et al. 2023), Big-Lama (Suvorov et al. 2022), Lama (Suvorov et al. 2022), and MADF (Zhu et al. 2021) are selected as the baselines in this evaluation. We select four canonical metrics, including structural similarity (SSIM), Peak Signal to Noise Ratio (PSNR), Mean Absolute Error (MAE), and Learned Perceptual Image Patch Similarity (LPIPS), to quantitatively measure the scores of all competitors.

Qualitative Comparison

Fig. 5 illustrates several inpainted images. Evidently, from the facial instances in the top two rows, it can be observed that most previous approaches, including Big-Lama, Lama, and MADF, produce visually defective results such as geometric distortions, blurred artifacts, incongruous eye colors, or identity disparities. AOT, Coordfill, and MAT perform relatively better. However, compared to these three, our SyFormer still stands out as producing the more visually authentic content, as can be easily witnessed from the hand segment. Furthermore, with respect to the scene images in the third row, the compared techniques again produce unwarranted textures at fence and ceiling positions, whereas our outcomes evince superior stability and competitiveness.

Quantitative Comparison

Table 1 presents the numerical results with different mask ratios, i.e., 30-40%, 40-50%, and 50-60%. As can be seen, compared to existing methods, our model consistently achieves superior or highly competitive performance across both datasets. Taking PSNR as an example, the average value of SyFormer from the two datasets is significantly better than AOT, Lama, MAT, MADF, Big-Lama, and Coordfill, with an improvement of 0.7587 dB, 1.0472 dB, 1.7247 dB, 1.3354 dB, 0.9868 dB, and 0.3258 dB, respectively.

Dataset		Places2		
Mask Ratio		30-40%	40-50%	50-60%
SSIM(↑)	w/o DRF	0.8308	0.7641	0.6414
	w/o FDA	0.8483	0.7845	0.6631
	w/o SCP	0.8361	0.7631	0.6489
	Full Model	**0.8552**	**0.7922**	**0.6695**
PSNR(↑)	w/o DRF	23.3944	21.7255	19.3818
	w/o FDA	24.0317	22.3508	19.8796
	w/o SCP	23.4699	21.8667	19.6183
	Full Model	**24.0885**	**22.3871**	**19.9352**
MAE(↓)	w/o DRF	0.0344	0.0471	0.0695
	w/o FDA	0.0323	0.0438	0.0643
	w/o SCP	0.0333	0.0462	0.0661
	Full Model	**0.0314**	**0.0431**	**0.0639**
LPIPS(↓)	w/o DRF	0.1372	0.1927	0.2831
	w/o FDA	0.1237	0.1727	0.2643
	w/o SCP	0.1369	0.1833	0.2812
	Full Model	**0.1161**	**0.1685**	**0.2423**

Table 2: Ablation study on Places2. The best values are highlighted by boldface and the second-best values are highlighted by underlines.

Ablation Experiments

Taking Places2 as an example, the roles of our proposals, including DRF, FDA, and SCP, are ablated in Table 2. The three labels, i.e., "w/o DRF", "w/o FDA", and "w/o SCP", in Table 2 denote the models with the concerned module removed and replaced with common treatments. Functionally, in comparison to the model "w/o DRF", the full version enjoys the ability to extract information from incomplete images and skills in global modeling of valid tokens. Similarly, the structural information captured by the SCP module is

Dataset		CelebA-HQ		
Mask Ratio		30-40%	40-50%	50-60%
SSIM(↑)	w/o DRF	0.9071	0.8615	0.776
	w/o FDA	0.9159	0.8773	0.7919
	w/o SCP	0.8888	0.8517	0.7667
	Full Model	**0.9184**	**0.8818**	**0.8001**
PSNR(↑)	w/o DRF	27.1105	25.1704	22.2953
	w/o FDA	27.8494	25.9098	22.9472
	w/o SCP	26.5621	24.9421	21.7414
	Full Model	**27.9467**	**26.0252**	**23.1334**
MAE(↓)	w/o DRF	0.0209	0.0307	0.0456
	w/o FDA	0.0188	0.0262	0.0417
	w/o SCP	0.0234	0.0326	0.0501
	Full Model	**0.0187**	**0.0259**	**0.0408**
LPIPS(↓)	w/o DRF	0.1036	0.1582	0.2234
	w/o FDA	0.0896	0.1284	0.1911
	w/o SCP	0.1094	0.1588	0.2251
	Full Model	**0.0763**	**0.1084**	**0.1640**

Table 3: Ablation study on CelebA-HQ. The best values are highlighted by boldface and the second-best values are highlighted by underlines.

beneficial for accurately matching pertinent patches, which relieves the shortcomings caused by the lack of image domain data. Besides, the FDA module is effective in combining multi-level information. As can be seen from Table 2, all these advantages are evidenced. Taking PSNR as an example, the gains of our SyFormer over "w/o DRF", "w/o FDA", and "w/o SCP" reach 0.6363 dB, 0.0495 dB, and 0.4853 dB, respectively.

Conclusion

In this paper, we present a new image inpainting network for large-portion missing case, namely the Structure-Guided Synergism Transformer (SyFormer), which successfully overcomes the challenges of modeling incoherent image distributions and assembling information from limited references. Technically, the proposal combines the power of pixel matching bootstrapping via structural priors with the robust long-range modeling via progressive filtering strategies. Concurrently, facilitated by a multi-level upsampling technique, SyFormer culminates in achieving high-quality inpainting results. Experimental evaluations on CelebA-HQ and Places2 datasets demonstrate the superior performance of SyFormer over state-of-the-art methods, both qualitatively and quantitatively.

Acknowledgements

This work was supported in part by the Pioneer and Leading Goose R&D Program of Zhejiang, under Grant 2023C01241, in part by the Key Program of Natural Science Foundation of Zhejiang, under Grant LZ24F030012, and in part by the National Natural Science Foundation of China under Grant 62276232.

References

Dong, Q.; Cao, C.; and Fu, Y. 2022. Incremental transformer structure enhanced image inpainting with masking positional encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11358–11368.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Feihong, L.; Hang, C.; Kang, L.; Qiliang, D.; Jian, Z.; Kaipeng, Z.; and Hong, H. 2023. Toward high-quality face-mask occluded restoration. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(1): 1–23.

Feng, Y.; Jiang, J.; Xu, H.; and Zheng, J. 2023. Change detection on remote sensing images using dual-branch multilevel intertemporal network. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–15.

Guo, X.; Yang, H.; and Huang, D. 2021. Image inpainting via conditional texture and structure dual generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 14134–14143.

Jin, Y.; Wu, J.; Wang, W.; Wang, Y.; Yang, X.; and Zheng, J. 2022. Dense vehicle counting estimation via a synergism attention network. *Electronics*, 11(22): 3792.

Jin, Y.; Wu, J.; Wang, W.; Yan, Y.; Jiang, J.; and Zheng, J. 2023. Cascading blend network for image inpainting. *ACM Transactions on Multimedia Computing Communications and Applications*. DOI: 10.1145/3608952.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*.

Li, J.; Wang, Z.; and Hu, X. 2021. Learning intact features by erasing-inpainting for few-shot classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 8401–8409.

Li, W.; Lin, Z.; Zhou, K.; Qi, L.; Wang, Y.; and Jia, J. 2022. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10758–10768.

Liao, L.; Xiao, J.; Wang, Z.; Lin, C.-W.; and Satoh, S. 2020. Guidance and evaluation: Semantic-aware image inpainting for mixed scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 683–700.

Liu, G.; Reda, F. A.; Shih, K. J.; Wang, T.-C.; Tao, A.; and Catanzaro, B. 2018. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 85–100.

Liu, W.; Cun, X.; Pun, C.-M.; Xia, M.; Zhang, Y.; and Wang, J. 2023. CoordFill: Efficient high-resolution image inpainting via parameterized coordinate querying. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1746–1754.

Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.; and Ebrahimi, M. 2019. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 3265–3274.

Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; and Efros, A. A. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2536–2544.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831.

Ren, Y.; Yu, X.; Zhang, R.; Li, T. H.; Liu, S.; and Li, G. 2019. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 181–190.

Song, L.; Cao, J.; Song, L.; Hu, Y.; and He, R. 2019. Geometry-aware face completion and editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2506–2513.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2149–2159.

Wan, Z.; Zhang, J.; Chen, D.; and Liao, J. 2021. High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4692–4701.

Xiong, W.; Yu, J.; Lin, Z.; Yang, J.; Lu, X.; Barnes, C.; and Luo, J. 2019. Foreground-aware image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5840–5848.

Xu, H.; Jiang, J.; Feng, Y.; Jin, Y.; and Zheng, J. 2022. Tensor completion via hybrid shallow-and-deep priors. *Applied Intelligence*, 1–22.

Xu, H.; Zheng, J.; Yao, X.; Feng, Y.; and Chen, S. 2021. Fast tensor nuclear norm for structured low-rank visual inpainting. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2): 538–552.

Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; and Huang, T. S. 2018. Generative image inpainting with contextual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5505–5514.

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

Zeng, Y.; Fu, J.; Chao, H.; and Guo, B. 2023. Aggregated contextual transformations for high-resolution image inpainting. *IEEE Transactions on Visualization and Computer Graphics*, 29(7): 3266–3280.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6): 1452–1464.

Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; and Lau, R. W. 2023. BiFormer: Vision transformer with bi-Level routing attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10323–10333.

Zhu, M.; He, D.; Li, X.; Li, C.; Li, F.; Liu, X.; Ding, E.; and Zhang, Z. 2021. Image inpainting by end-to-end cascaded refinement with mask awareness. *IEEE Transactions on Image Processing*, 30: 4855–4866.

Zuo, Z.; Zhao, L.; Li, A.; Wang, Z.; Zhang, Z.; Chen, J.; Xing, W.; and Lu, D. 2023. Generative image inpainting with segmentation confusion adversarial training and contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3888–3896.