# Multi-scale Gated Inpainting Network with Patch-Wise Spacial Attention

Xinrong Hu[1,2], Junjie Jin[1,2(✉)], Mingfu Xiong[1,2], Junping Liu[1,2], Tao Peng[1,2], Zili Zhang[1,2], Jia Chen[1,2], Ruhan He[1,2], and Xiao Qin[3]

[1] Engineering Research Center of Hubei Province for Clothing Information, Wuhan, China
junjie.jin@qq.com
[2] School of Mathematics and Computer Science, Wuhan Textile University, Wuhan, China
[3] Department of Computer Science and Software Engineering, Aubern University, Auburn, USA

**Abstract.** Recently, deep-model-based image inpainting methods have achieved promising results in the realm of image processing. However, the existing methods produce fuzzy textures and distorted structures due to ignoring the semantic relevance and feature continuity of the holes region. To address this challenge, we propose a detailed depth generation model (GS-Net) equipped with a Multi-Scale Gated Holes Feature Inpainting module (MG) and a Patch-wise Spacial Attention module (PSA). Initially, the MG module fills the hole area globally and concatenates to the input feature map. Then, the module utilizes a multi-scale gated strategy to adaptively guide the information propagation at different scales. We further design the PSA module, which optimizes the local feature mapping relations step by step to clarify the image texture information. Not only preserving the semantic correlation among the features of the holes, the methods can also effectively predict the missing part of the holes while keeping the global style consistency. Finally, we extend the spatially discounted weight to the irregular holes and assign higher weights to the spatial points near the effective areas to strengthen the constraint on the hole center. The extensive experimental results on Places2 and CelebA have revealed the superiority of the proposed approaches.

**Keywords:** Image inpainting · Feature reconstruction · Gated mechanism · Spacial attention · Semantic relevance

## 1 Introduction

The goal of image completion is the task to fill the missing pixels in an image in a way that the corresponding restored image to have a sense of visual reality. The restored area needs continuity and consistency of texture while seeking semantic consistency between the filled area and any surrounding area. Image completion techniques are widely adopted in photo recovery, image editing, object deletion and other image tasks [1, 5]. At present, the existing methods have focused on the restoration of the rectangular areas near the image centers [6, 7]. This kind of regular hole restoration could result in the model over-fitting accompanied by poor migration effect [8]. The overarching objective of this


C. S. Jensen et al. (Eds.): DASFAA 2021 Workshops, LNCS 12680, pp. 169–184, 2021.
https://doi.org/10.1007/978-3-030-73216-5_12

work is to propose an image-restoration model, which is sufficiently robust to repair regular and irregular holes. Our proposed technique produces semantically meaningful predictions to ensure that the repaired parts are perfectly integrated with other portions without any expensive post-processing.

Traditional image restoration methods mainly exploit the texture synthesis technology to address the challenge of hole fillings. These methods assume that the missing regions should contain a pattern similarity to those of background regions. And they use the certain statistics of the remaining image to restore the damaged image region [1–4]. As one of the most advanced techniques used in the past, PatchMatch [1] can quickly find the nearest neighbor matching to replace the repaired hole area through the stochastic algorithm. Although it usually produces the smooth results especially in background rendering tasks, it is limited by the available image statistics and just considers the low-level structures without any high-level semantics or global structures for captured images. In addition, the traditional diffusion-based and block-based methods assume that missing blocks can be found in the background image and they cannot generate new image content for complex and non-repetitive structural regions (e.g. human faces) [9].

Nowadays, the deep-learning-based methods are constantly explored to overcome the aforementioned obstacles of the above methods by training a large amount of data [8–10, 12, 13]. In particular, deep convolutional neural networks (CNNs) and generative adversarial networks (GANs) have been introduced to implement the image complement tasks [9, 14, 15]. Broadly speaking, image inpainting tasks equipped with deep module mainly can be divided into two categories. The first ones uses global spatial attention to fill holes by building the similarity between a missing area and the other areas [6, 7, 19, 31]. Although this group of methods can ensure a consistency between generated information and context semantics, there often exist pixel discontinuity and semantic gaps [12]. The second family of schemes is to attach different levels of importance to the valid pixels of the original image to predict the missing pixels [8, 14]. These methods correctly handle irregular vulnerabilities correctly, but the generated content still suffers from semantic errors and boundary artifacts [12]. The above methods work poorly due to ignoring the semantic relevance and the feature continuity of the generated contents, which are related to the continuity of local pixels.

Inspired by the human mind coupled with the partial convolution [8], we propose a Multi-Scale Gated Inpainting module (MG) and a Patch-wise Spacial Attention module (PSA) are proposed to fill an unknown area of the feature map with similar method as a part of our model. The MG module first fills each unknown feature patch in an unknown region with the most similar feature patch in the known regions. Subsequently, the selection of information in the filled area is controlled by a two-scale gating strategy. As a result, the global semantic consistency is guaranteed by the first step, and the local feature consistency is optimized by the second step optimization. In addition to controlling the style relation under local features, the PSA module handles the repaired features with block-level attention.

Technically, our model uses the U-Net [20] architecture as a baseline to propagate the consistency of global and local styles and detailed texture information to the missing areas. On the whole, this model continuously collects the features of an effective region

through partial convolution. At a higher level of the encoding phase, we develop a distinctive Multi-Scale Gated Inpainting module (MG) to carry out two phases. First, MG revises a current hole area for its global style alignment through Contextual Attention [6]. Second, MG brings a resulting feature fix into alignment with the overall style through a multi-scale gated mechanism. In the decoding stage, PSA divides the feature channel into multiple patch blocks to further optimize the consistency of local styles so that the network learns more effective local features. Finally, the repaired image is delivered to VGG16 [29] to gauge the style loss and perception loss. Which help generate details consistent with the global style. In addition, our model is finally down-sampled to the size of $4 \times 4$ in order to obtain a higher level of semantic consistency. The experiments driven by the two standard datasets (Places2 [25] and CelebA [26]) reveal that the proposed methods produce higher quality results than those of the existing competitors. The main contributions of this work are summarized as follows:

- We develop a new Multi-Scale Gated Inpainting module (MG) applied to the model structure. MG combines feature maps generated by gated modules of different proportions to obtain structural information of features at different scales, thereby flexibly leveraging background information to balance the image requirements.
- We extend the spatial attention module by adding the minimization feature of patch-wise to ensure that the pixels generating holes area are true and locally stylized.
- We introduce the concepts of style loss and the perception loss to construct the proposed loss function, which yield a consistent style. The proposed new spatial discounted loss of irregular holes helps to strengthen hole-center constraints, thus promoting texture consistency.
- The experiments with two standard datasets (Places2 [25] and CelebA [26]) demonstrate the superiority of our approaches over the most advanced methods found in the literature.

## 2 Related Work

### 2.1 Image Inpainting

Traditional non-learning methods propagate and reproduce information by calculating the similarity with the other background regions [2, 4]. PatchMatch [1] can well synthesize surface textures through the nearest neighbor matching algorithm, which is an excellent patch matching algorithm. However, these methods do not semantically originate meaningful contents, neither can the methods deal with large missing areas. For the nonexistent detailed texture features, these schemes are unable to generate new features while exhibiting poor recovery effect.

In recent years, the methods based on deep learning have become a significant symbol of the image restoration. Context Encoder [15] tries to restore the central area ($64 \times 64$) of $128 \times 128$ images. This technique is the first deep network model to handle the inpainting tasks, which provides reasonable results for the holes semantic filling. Unfortunately, it has a poor inpainting ability at fine textures. Shortly thereafter, Iizuka *et al.* extends the context encoder by proposing local and global discriminators to improve repaired quality for the image consistency [10]. This extension overlooks the consistent relation

between holes and the other areas as a whole. Therefore, there exist more obvious color differences. The Context Encoder [15] is trained to act as the constraint of global content [13], local texture constraints are constructed by using the local patches similarity of the missing part and the known regions to obtain high-resolution prediction.

### 2.2 Feature Matching and Feature Selection

Global spatial attention mechanisms have also been deployed to address image inpainting challenges by the virtue of similarity relation. CA [6] creates a rough prediction for the hole area through similarity calculation. The Multi-Scale Contextual-Attention [7] patch is located in the missing region. Ultimately, the re-weight for both is located on the Squeeze-And-Excitation [16] module, which improves generalization ability of the model. SCA [12] build a patch-wise continuity relationship between adjacent features of the missing area to enhance the continuity of features inside the holes area. Shift-net [19] selects a specific encode-decode layer of the same level for similarity measures, encoder features of the known region are shifted to serve as an estimation of the missing parts. The RFR [31] harvests remote information and progressively infers a boundary of the hole by the KCA module, thereby gradually strengthening the constraints on the hole's center.

The above methods adopt the similar treatment for the corruption areas and non-corruption area, thereby leading to artifacts such as color discrepancy and blurriness. Only the effective features of each layer are processed by partial convolution [8]. By updating the mask of each layer and normalizing the convolution weights and mask values, which ensures that the convolution filter focuses on the effective information of the known regions to deal with irregular holes. Partial convolution is regarded as a kind of hard mask [14], which confronts roadblocks learn specific mask information. Furthermore, it introduces automatic learning soft mask by using gated convolution and combines with SN-Patch GAN discriminator to achieve optimized predictions. When it comes to feature normalization, the above methods do not consider the influence of mask areas, which limits the trainings of network repair. Treating the damaged areas and the undamaged areas separately [11], the mean value and variance deviation are solved to continuously improve network performance.

Unlike the leading-edge strategies proposed in the literature [6, 8, 14], our solution is tailored for process images where backgrounds are misleading or lacks similarity. Our technique has an edge over the existing methods, because ours leverages the multi-scale gated module to control the degree of feature extraction while dynamically screening useful features to alleviate the problem of information redundancy and loss. In order to enrich repaired details, an extended spatial attention module [28] performs the patch-wise division on the channel to dynamically extract local features. In this way, our model is adept at generalizing scenes and understanding styles as well as picture details.

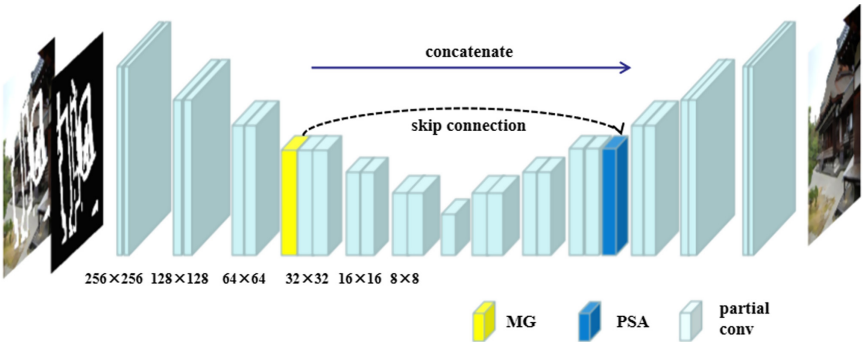



**Fig. 1.** The architecture of our GS-Net model. We augment the MG and PSA layer at the resolution of 32 × 32 in the network.

## 3 Approach

We describe the entire model structure from top to bottom, and then introduce the MG module and PSA module in details. Some extensions to the model are also expressed to allow optional user guidance.

### 3.1 An Overview of Our Model (GS-Net)

Our model is a one-stage and end-to-end image inpainting model, thereby making our approach simpler and easier to implement than other methods. More specifically, U-Net [20] is used as the baseline structure and the partial convolution [8] is stacked as the basic modules for deep feature extraction in GS-Net (see also Fig. 1). More formally, we denote W as convolution layer filter weights, b as the bias, M as the mask, and X as the feature values for the current convolution window. The partial convolution is expressed as:

$$x^{'} = \begin{cases} W^{T}(X \otimes M)\frac{sum(1)}{sum(M)} + b, & if\ sum(M) > 0 \\ 0, & otherwise \end{cases} \tag{1}$$

In each convolution window with effective feature values, partial convolution layer assigns greater weight to the convolution result with fewer feature values through the above operation. After each partial convolution operation is accomplished, whether the mask has a valid pixel update mask through the convolution region. This process is expressed as:

$$m^{'} = \begin{cases} 1, & if\ sum(M) > 0 \\ 0, & otherwise \end{cases} \tag{2}$$

After the feature map passes through partial convolutional layer, the missing area is filled with the surrounding effective feature area and becomes smaller. Therefore, all the features areas of the holes will be completely filled after sufficient successive applications of the partial convolution layer.

### 3.2 Multi-scale Gated Inpainting Module (MG)

Partial convolution in our model is stacked with layers to update masks and feature maps. In partial convolution, the holes region gradually disappear with the deepening of convolution depth, which is conducive to extracting effective depth features. However, directly interpolating the features of empty regions from the features of non-empty regions during an up-sampling process leads to the final blurry texture of missing regions. The root of such a problem is to directly extract features without repairing the features of holes, thereby ignoring the spatial continuity of features.

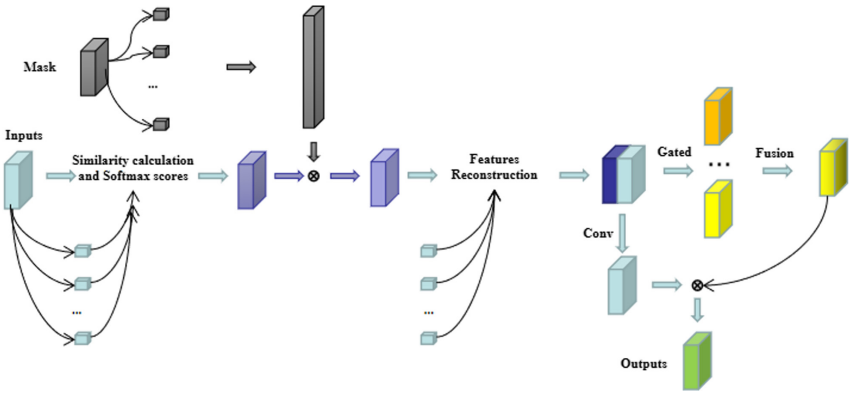



**Fig. 2.** In the MG Module, an input feature is transferred to a global attention module to fill a hole feature area, which is concatenated into the input feature to obtain different gated scores through two convolution kernels of multiple sizes ($3 \times 3$ and $1 \times 1$). In the end, element-wise multiplications are performed specifically by multiplying the convoluted concatenate by the two gated scores of fusion as an output feature map.

To address the challenge of blurred image content and distorted structure, we propose a repair network with MG module in Fig. 2. First of all, The MG module uses CA [6] method to fill the holes in the high-level feature. Partial convolution is still the interpolation of depth features on the hole area, and the inability to match the optimal patch leads to information loss and confusion. Therefore, CA algorithm is used to construct similarity matrix, and deconvolution operation ensures the trainability of interpolation process.

Given an input feature map $\phi_{in}$, we firstly replace fore-ground feature map using attention mechanism. For each attention map, we use similar strategy to calculate scores as [6] the calculation of attention score could be implemented as convolution calculation.

$$s_{x,y,x',y'} = < \frac{f_{x,y}}{||f_{x,y}||}, \frac{b_{x',y'}}{||b_{x',y'}||} > \tag{3}$$

where, $f_{x,y}$, $b_{x',y'}$ are fore-ground patches and background patches respectively. $s_{x,y,x',y'}$ is similarity matrix between all the patches. To compute the weight of each patch, softmax is applied on the channel of score map to obtain softmax scores. Since any change in

the foreground patch is more related to the similar change in the background patch. CA adopts a left-right propagation followed by a top-down propagation with kernel size of k, and then propagate score to better merge patch.

$$\hat{s}_{x,y,x',y'} = \sum_{i\in\{-k,\ldots,k\}} s^*_{x+i,y,x'+i,y'} \tag{4}$$

where $s^*$ is channel-level softmax applied to feature mapping. Finally, the multichannel mask multiplication is used to preserve the current information and then deconvolution operation are responsible for restoring a missing feature map. However, the misleading highly similar regions and the existence of the hole regions could lead to the disappearance of effective features in deconvolution, which is detrimental to feature restoration. Fortunately, we are inspired by gated convolution to ensure the dynamic selection of effective features by constructing a soft mask mechanism, in which the expected output features are learned under fixed gateway scale. This learning mechanism is dynamic, general and accurate. When the hole feature is repaired, other inappropriate features are incorrectly filled. And the gate control mechanism dynamically adjusts the gate value to construct an appropriate output feature.

Moreover, it is nontrivial to determine the appropriate patch match size for various image to reveal images details. In general, larger patch size helps ensure style consistency while smaller patch size is more flexible on using background feature map. Patch matching on a single fixed scale seriously limits the capability to fit the model into different scene [7]. To this end, we devise a novel MG module that helps to make use of background content flexibly based on the overall image style. The MG integrate feature selection at two different scales by convolution opteration. In order to better distinguish the importance of the two, we simply use a learnable parameter $\lambda$ as the dynamic threshold. Formally, the gated feature of the output is written as:

$$output_{gated} = \lambda output_1 + (1-\lambda) output_2 \tag{5}$$

where, $output_1$ and $output_2$ are two gated values at two different scales. Therefore, the value output by the MG module has comprehensive information at multiple scales.

Through the MG module's information, global features become continuous thanks to global-feature-hole fillings and the multi-scale gated selection mechanism. Therefore, the partial convolution layer has no need to distinguish between the region of holes and non-holes and; thus, the mask is set to 1. The MG module is adroit at capturing background information on high-level semantics while producing contents with elegant details.

### 3.3 Patch-Wise Spacial Attention Module (PSA)

A vital feature of a human visual system is that people have no intent to process an entire scene at once. Instead, humans take advantage of a series of local glimpses and selectively focus on salient parts in order to capture visual structure in a swift manner. Although the attention mechanism is widely used in image classification, the mechanism has no appearance in image inpainting. An important reason is that when it comes to

incomplete image, there may still be hole information in high-level features. At this time, the traditional s spatial attention mechanism gives rise to structure dispersion and texture loss in generated images. In order to solve the above challenges and take advantage of the attention mechanism, we extend the CBAM technique [28] to devise a patch-wise attention mechanism depicted in Fig. 3.

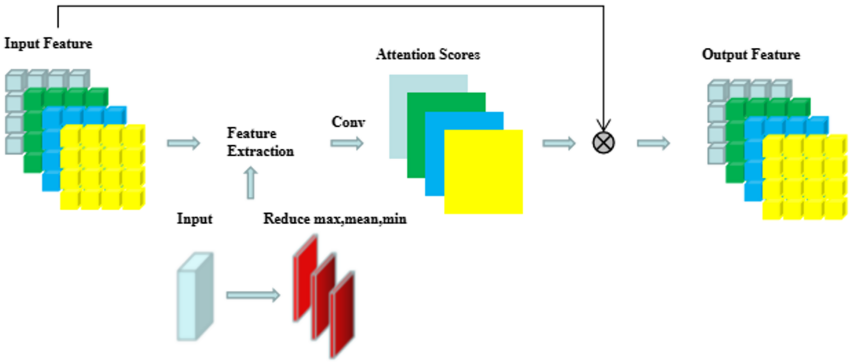



**Fig. 3.** PSA Module. 3D features are divided into 3 × 3 × 4 small blocks, and the maximum, average and minimum values of channel-wise are calculated for each small block, and the spatial attention of each block is obtained through convolution.

Instead of using the channel-wise reduction directly, we choose a 3 × 3 × 4 feature block as a unit of attention restoration in order to ensure the consistency relationship of local features and maximize the effect of the restoration of the hole area. We first apply maximum pool, average pool, and minimum pool operations on the channel axis and concatenate them to produce a valid feature descriptor. On the feature descriptor of the connection, convolution layer is applied to generate the spatial attention graph and extract the information features of each patch through the convolution operation. Different from classification and identification tasks, the minimum pool operation obtains the characteristics of possible hole repair to hold hole is emphasized so as to ensure its information flow in the network.

We aggregate channel information of a feature map by using three pooling operations, generating three 2D maps in $i^{th}$ patch (The $i^{th}$ patch here represents the channel between (i − 1) * 4 and i * 4): $F^i_{\max}, F^i_{avg}, F^i_{\min} \in R^{1\times H\times W}$. Each denotes max-pooled features, average-pooled features and min-pooled features across the channel. Those features are then concatenated and convolved by a standard convolution layer, producing our 2D patch-wise spatial attention map. In short, the spatial attention is computed as:

$$M_{\mathrm{i}}(F) = \sigma(f^{3\times 3}([F^i_{\max}, F^i_{avg}, F^i_{\min}])) \tag{6}$$

where $\sigma$ denotes the sigmoid function and $f^{3\times 3}$ represents a convolution operation with the filter size of 3 × 3.

## 4 Loss Function

Similar to the design of PC [8], the style consistency and detail level are also taken into the consideration of our loss function. For the model learning process can fully pay attention to the texture details and structural information, we consider a pre-trained VGG16 [29] as a fixed model to extract high-level features. The perceptual loss [30] and style loss compare the difference between the deep feature map of the generated image and the ground truth under different descriptors. All parameter symbols are described as follows. $\phi_i$ denotes feature maps $i^{th}$ pooling layer. H, W, C refer to the height, weight and channels number for a feature map, respectively. And N is the number of feature maps generated by the VGG16 feature extractor. The perceptual loss can be expressed as follows:

$$L_{perceptual} = \sum_{i=1}^{N} \frac{1}{H_i W_i C_i} |\phi_i^{gt} - \phi_i^{out}|_1 \tag{7}$$

Although perceptual loss helps to capture high level structures, the perceptual loss lacks the ability to preserve style consistency. To address this drawback, we advocate for the style loss ($L_{style}$) as an integral apart of our loss function. With the help of the style loss, our model is adroit at learning color and overall style information from backgrounds.

$$L_{style} = \sum_{i=1}^{N} |\frac{1}{H_i W_i C_i} (\phi_i^{style_{gt}} - \phi_i^{style_{out}})|_1 \tag{8}$$

$$\phi_i^{style} = \phi_i \phi_i^T \tag{9}$$

Total variation (TV) loss $L_{tv}$, the smoothing penalty [30] on R, is introduced into the loss function. Here R is the area of a sliding window that contains missing pixels. However, the cost of directly applying TV losses to a hole area is to promote texture blurring of the hole area. More unfortunately, in the case of large losing areas, this approach leads to a failure to repair void areas -- the hole areas remain void areas. In order to address the problem of huge amount, we benefit from two cognitions: a hole area has a certain similarity with the TV of surrounding areas; the edge of the hole area maintains a certain continuity with the surrounding area. The TV loss is expressed as follows:

$$L_{tv} = L_{row} + L_{col} \tag{10}$$

$$L_{row} = \sum_{(i,j)\in R,(i,j+1)\in R} \frac{||I_R^{i,j+1} - I_R^{i,j}||_1}{N_{I_R}}, \; L_{col} = \sum_{(i,j)\in R,(i+1,j)\in R} \frac{||I_R^{i+1,j} - I_R^{i,j}||_1}{N_{I_R}} \tag{11}$$

where $I_R^{i,j}$ represents an image pixel point, $N_{I_R}$ is defined as the number of elements in the hole's region. Especially for large holes, boundaries are sometimes still artifacts, which may be the lack of constraints on the center of the holes. Similar to the spatial

discounted loss of the CA algorithm, the closer the hole region is to the known region, the more attention should be given to it. However, the distance between a hole point and a surrounding effective region is difficult to calculate when it comes to irregular holes. To simplify the computation, we traverse the symmetric mask value near each hole point and undertake a bit operation to quickly obtain the hole length. This process is formally articulated as follows:

$$Sym_{i,j,length} = mask_{i,j}^{i-length,j} | mask_{i,j}^{i+length,j} \tag{12}$$

where $mask_{i,j}^{i-length,j}$, $mask_{i,j}^{i+length,j}$ are the fields of length at the coordinates of point (i, j). Finally, our target becomes the maximum length value of Sym = [0].

$$discounted_{left-right} = \max_{length} \{Sym_{i,j,length} = [0]\} \tag{13}$$

where [0] represents the matrix with all values of 0. The upper and lower relation of the hole region is solved by the same strategy and denoted as $discounted_{top-bottom}$. Therefore, the total spatial discounted weight is formalized as follows:

$$discounted = \gamma^{(discounted_{left-right} + discounted_{top-bottom})/2} \tag{14}$$

where, $\gamma$ represents a weighting factor. Futher, $L_{valid}$ and $L_{hole}$ which calculate L1 differences in the unmasked area and masked area respectively. The total loss $L_{total}$ is the combination of all the above loss functions. Thus, we have

$$L_{total} = \lambda_{vaild} L_{valid} + \lambda_{hole}(L_{hole} \odot discounted) + \lambda_{perceptual} L_{[erceptual} + \lambda_{style} L_{style} + \lambda_{tv} L_{tv} \tag{15}$$

## 5 Experiments

### 5.1 Datasets and Experimental Details

In this section, we evaluate our model on two datasets: the Places2 [25] dataset and the CelebA [26] dataset. The Places2 dataset is a garden scene selected from the Places365-Standard dataset, which embraces 9069 images. The dataset is divided into the train, validate, and test subsets with a ratio of 8:1:1. The CelebA dataset contains 162,770 training images, 19,867 validation images, and 19,962 test images. We use both the training set and validation set for training purpose, whereas the test set is dedicated for testing. In the end, we use the mask dataset of partial convolution, which contains 55,116 masks for the training and 24,866 masks for testing. The size of these masks is 512 × 512. After resizing these masks to 256 × 256, we place the masks into our network model.

For all the parameter settings similar to those elaborated in the literature [8], the tradeoff parameters are set as $\lambda_{valid} = 1$, $\lambda_{\text{hole}} = 6$, $\lambda_{perceptual} = 0.05$, $\lambda_{style} = 120$ and $\lambda_{tv} = 0.1$. Our model is initialized the weights using the initialization method described in [9] and use Adam [27] for optimization with a learning rate of 0.0001, and train on a single NVIDIA V100 GPU (32 GB) with a batch size of 6. The Places2 models are

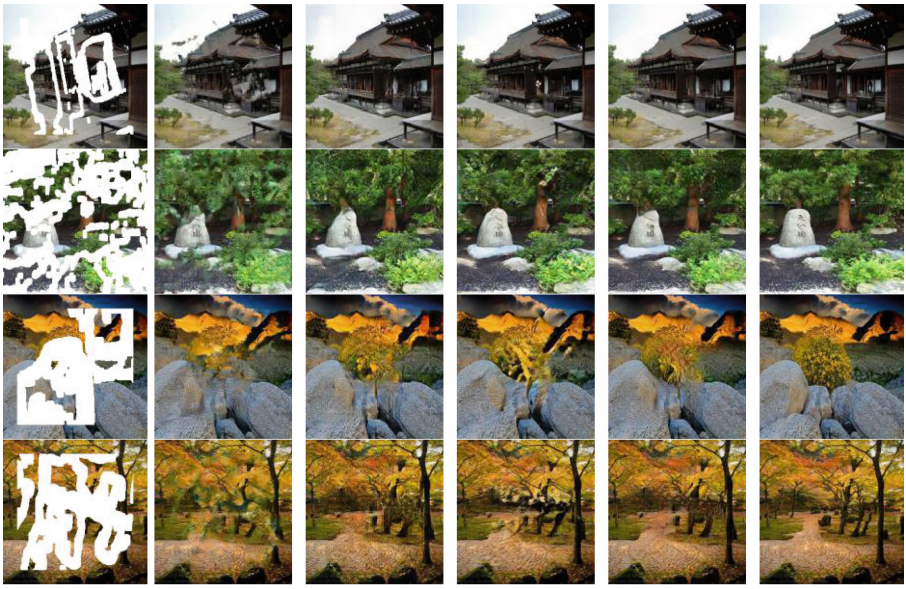



**Fig. 4.** A Comparison of test results on Places2 images.

trained for two days, whereas the CelebA models are trained for approximately one week.

We compare the proposed MG and PSA algorithm with the following three state-of-the-art methods: CA [6]: Contextual Attention, PC [8]: Partial Convolution, GC [14]: Gated Convolution.

To make fair comparisons with the CA and GC approaches, we retrain the CA and GC models on the same datasets. Both CA and GC methods are trained using a local discriminator available in a local boundary box of the hypothetical hole, which makes no sense for the shape of masks [8]. As such, we directly use CA and GC released pre-trained models. And PC is trained under the same conditions as those in our experimental setup until the PC model is converged.

### 5.2 Qualitative Comparisons

Figure 4 unveils the comparison results among our method and the three most advanced approaches processing in the Places2 dataset. All images are displayed at the same resolution (256 $\times$ 256). The CA approach is effective at semantic inpainting, but the results shown above appear to be abnormally blurry and artifact. The PC method fills the hole areas with the corresponding styles, but PC loses some of the detail textures. The GC method exhibits a strong inpainting ability in local details and overall styles. Unfortunately, GC suffers from the local overshine problems. Compared with the other methods, our solution has an edge under large hole conditions by originating inpainting results that alleviate artificial traces. Figure 5 unravels that our model is able to generate fully detailed, semantically plausible, and authentic images with superb performance.


**Fig. 5.** Comparison of test results on CelebA images.

### 5.3 Quantitative Comparisons

Now we quantitatively evaluate our model on the two datasets, using three quality methods, namely, the structural similarity (SSIM), peak signal-to-noise ratio (PSNR), and mean L1 loss assessment image similarity. Because the image restoration application in the application scenario will not stick to the above mask structure. To make a fair numerical comparison, we apply the mask generation method of GC [14] to compare the mask repair effects under the three different proportions in Fig. 6. One thousand masks and their corresponding random pictures are elected in the tests, the results of which are recapped in Table 1.


**Fig. 6.** Some test masks for each hole-to-image area ratio category.

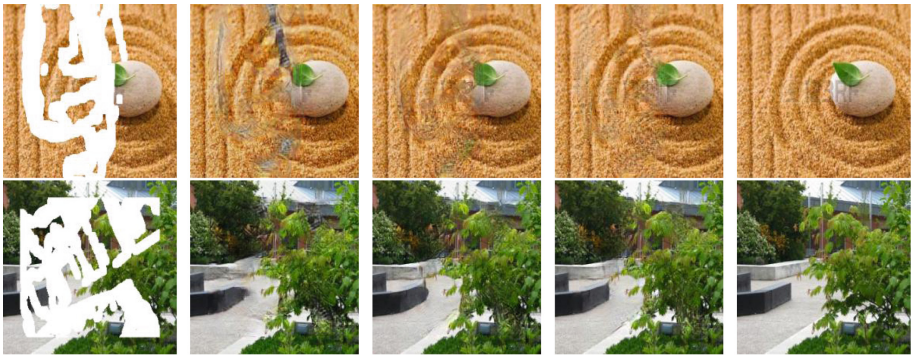
Table 1 illustrates that our method produces the decent results with the best SSIM, PSNR and mean l1 loss on the Places2 dataset and the CelebA faces dataset. Similar to the aforementioned results, our MG and PSA algorithm is a front runner in terms of numerical performance on the Places2 and CelebA datasets. When it comes to repairing large holes, the performance improves of our algorithm over the existing techniques become more pronounced.

**Table 1.** Numerical comparison on two datasets.

| Dataset | | Places2 | | | CelebA | | |
|---|---|---|---|---|---|---|---|
| Mask ratio | | 10%–20% | 30%–40% | 50%–60% | 10%–20% | 30%–40% | 50%–60% |
| Mean l1(%) | CA | 3.0941 | 5.7280 | 7.4529 | 3.2433 | 6.0052 | 8.4367 |
| | PC | 3.2495 | 4.4537 | 5.3843 | 1.8712 | 2.5208 | 3.2301 |
| | GC | 2.0385 | 3.5036 | 4.7996 | 1.2488 | 2.1232 | 2.9248 |
| | Ours | 2.1274 | 3.1917 | 4.2045 | 0.9542 | 1.5228 | 2.0834 |
| PSNR | CA | 21.5031 | 18.1033 | 17.2827 | 20.8873 | 17.5012 | 16.0160 |
| | PC | 24.7846 | 22.1610 | 21.1155 | 29.3626 | 26.7636 | 24.9999 |
| | GC | 24.7426 | 21.5232 | 20.1670 | 28.6721 | 25.5052 | 23.9649 |
| | Ours | 25.7142 | 23.1374 | 22.0227 | 32.0948 | 28.9088 | 27.0451 |
| SSIM | CA | 0.8327 | 0.7042 | 0.6080 | 0.8337 | 0.7067 | 0.6015 |
| | PC | 0.8296 | 0.7307 | 0.6476 | 0.9050 | 0.8567 | 0.8074 |
| | GC | 0.8638 | 0.7623 | 0.6758 | 0.9180 | 0.8589 | 0.8050 |
| | Ours | 0.8650 | 0.7762 | 0.6917 | 0.9463 | 0.9061 | 0.8638 |

### 5.4 Ablation Study and Discussion

GS-Net, being carried out on partial convolution, is equivalent to the superposition processing of partial convolution layer excluding our proposed two modules. To clearly present the effectiveness of these operations, we compare various indicators by respectively removing the MG and PSA modules in Places2 dataset. Figure 7 and Table 2 reveal that compared to the results yielded by our algorithm, the results from the non-MG and non-PSA models exhibit more artifacts and distortions. At the same time, the MG module is superior to the PSA module in terms of performance index.



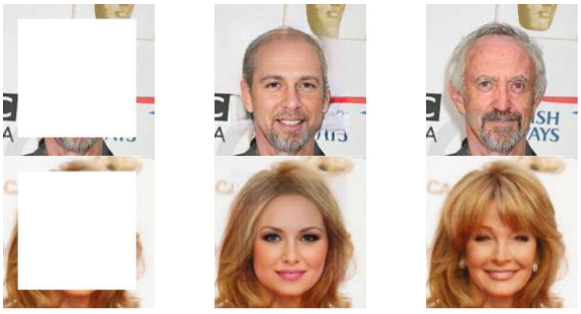
(a) Input (b) without MG (c) without PSA (d) ours (e) GT

**Fig. 7.** Comparison results for different attention manners. From the left to the right are: (a) Input, (b) Without MG, (c) Without PSA, (d) MG + PSA, (e) Ground Truth

**Table 2.** Numerical comparison on Places2 dataset.

| Method | Mean l1 loss (%) | PSNR | SSIM |
|---|---|---|---|
| Without MG | 3.8346 | 21.2940 | 0.7283 |
| Without PSA | 3.5739 | 22.5376 | 0.7615 |
| With MG and PSA | 3.1411 | 23.2501 | 0.7808 |

Apart from delivering strong capabilities in terms of recovery, GS-Net can be widely applied to intelligent face modification or face synthesis. Figure 8 shows two faces with different detail textures.



Input Ours GT

**Fig. 8.** In face of effect.

Features will be input into the PSA module after a globally filled hole area of the MG module. The information flowing through the MG module is well repaired, this PSA module is focused on controlling the relationship among local feature blocks. The PSA is constructed by the channel-wise attentional processing of local 3D blocks, thereby forming local relations such as local maximum, average, and minimum. It is evident that each patch repaired may be larger than unrepaired feature values. Thus, exerting an attention will pay more attention to the repaired local 3D region features, which is beneficial to the subsequent upsampling process.

Our model outperforms the cutting-edge techniques in most tested cases, but the repair effect still has a certain difference under a pure color background. The reason may be caused by partial convolution, which will be addressed in our foreseeable future research pathway.

## 6 Conclusion

We proposed in this paper the MG module, which is capable of gradually enriching the information of mask regions by offering semantically consistent embedding results . We developed the PSA module to further promote the enrichment of local texture details.

We conducted extensive qualitative and quantitative comparisons against the leading-edge solutions. The validity analysis and ablation learning demonstrate that our GS-Net outperforms the existing solutions over the Places2 and CelebA datasets.

## References


1. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: PatchMatch: a randomized correspondence algorithm for structural image editing. TOG **28**(3), 24:1–24:11 (2009)
2. Ballester, C., Bertalmio, M., Caselles, V., Sapiro, G., Verdera, J.: Filling-in by joint interpolation of vector fields and gray levels. IEEE Trans. Image Process. **10**(8), 1200–1211 (2018)
3. Criminisi, A., Pérez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE TIP **13**(9), 1200–1212 (2004)
4. Wilczkowiak, M., Brostow, G. J., Tordoff, B., Cipolla, R.: Hole filling through photomontage. In: Proceedings of the British Machine Vision Conference (BMVC), pp. 492–501. British Machine Vision Association, Oxford (2005)
5. Shetty, R., Fritz, M., Schiele, B.: Adversarial scene editing: automatic object removal from weak supervision. In: Thirty-second Conference on Neural Information Processing Systems, pp. 7717–7727. Curran Associates, Montréal Canada (2018)
6. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T. S.: Generative image inpainting with contextual attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5505–5514 (2018)
7. Wang, N., Li, J., Zhang, L., Du, B.: Musical: multi-scale image contextual attention learning for inpainting. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI), pp. 3748–3754 (2019)
8. Liu, G., Reda, F.A., Shih, K.J., Wang, T.-C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11215, pp. 89–105. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01252-6_6
9. Zhou, T., Ding, C., Lin, S., Wang, X., Tao, D.: Learning oracle attention for high-fidelity face completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7680–7689 (2020)
10. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM TOG **36**(4), 1–4 (2017)
11. Yu, T., et al.: Region normalization for image inpainting. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 12733–12740 (2020)
12. Liu, H., Jiang, B., Xiao, Y., Yang, C.: Coherent semantic attention for image inpainting. In: ICCV, pp. 4170–4179 (2019)
13. Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H.: High-resolution image inpainting using multi-scale neural patch synthesis. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6721–6729 (2017)
14. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In Proceedings of ICCV, pp. 4471–4480 (2019)
15. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: feature learning by inpainting. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2536–2544 (2016)
16. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141 (2018)

17. Nazeri, K., Ng, E., Joseph, T., Qureshi, F., Ebrahimi, M.: Edgeconnect: structure guided image inpainting using edge prediction. In Proceedings of ICCV Workshops (2019)
18. Xiong, W., et al.: Foreground-aware image inpainting. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5840–5848 (2019)
19. Yan, Z., Li, X., Li, M., Zuo, W., Shan, S.: Shift-net: image inpinting via deep feature rearrangement. In: Proceedings of ECCV, pp. 3–19 (2018)
20. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
21. Levin, A., Zomet, A., Weiss, Y.: Learning how to inpaint from global image statistics. In: Proceedings of International Conference on Computer Vision (ICCV), pp. 305–312 (2003)
22. Ding, D., Ram, S., Rodríguez, J.J.: Image inpainting using nonlocal texture matching and nonlinear filtering. IEEE Trans. Image Process. **28**(4), 1705–1719 (2018)
23. Snelgrove, X.: High-resolution multi-scale neural texture synthesis. In: SIGGRAPH Asia Technical Briefs, pp. 1–4 (2017)
24. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M.: Generative adversarial networks. In: NIPS, pp. 2672–2680 (2014)
25. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A10 million image database for scene recognition. IEEE TPAMI **40**(6), 1452–1464 (2018)
26. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3730–3738 (2014)
27. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
28. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
30. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
31. Li, J., Wang, N., Zhang, L., Du, B., Tao, D.: Recurrent feature reasoning for image inpainting. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7760–7768 (2020)
32. Zheng, C., Cham, T. J., Cai, J.: Pluralistic image completion. In: CVPR, pp. 1438–1447 (2019)