SA-RFR: self-attention based recurrent feature reasoning for image inpainting with large missing area

Human-centric Computing and Information Sciences

July 2022 | Volume 12

www.hcisjournal.com

KIPS Korea Information Processing Society

KIPS CSWRG Korea Information Processing Society Computer Software Research Group

HCIS | Human-centric Computing and Information Sciences

RESEARCH (Research Manuscript) Open Access

Human-centric Computing and Information Sciences (2022) 12:31
DOI: https://doi.org/10.22967/HCIS.2022.12.031
Received: September 28, 2021; Accepted: January 16, 2022; Published: July 15, 2022

SA-RFR: Self-Attention Based Recurrent Feature Reasoning for Image Inpainting with Large Missing Area

Jin Wang[1], Liu Wang[1], Shiming He[1,*], Osama Alfarraj[2], Amr Tolba[2], and R. Simon Sherratt[3]

Abstract

With the recent emergence of artificial intelligence, deep learning image inpainting methods have achieved fruitful results. These methods generated plausible structures and textures in repairing images with small missing areas. When inpainting an image with an excessively large missing area (the mask ratio is more than 50%), however, it usually produces a distorted structure or a fuzzy texture that is inconsistent with the surrounding area. Therefore, we propose a self-attention based recurrent feature reasoning (SA-RFR) network. First, SA-RFR uses self-attention (SA) to enhance the correlation between known pixels and unknown pixels and the constraints on the hole center, so that the repaired content details are clearer and the edges are smoother. In addition, because ordinary convolution has feature redundancy for the generated feature map, some unnecessary information is generated, and some models are difficult to train. Therefore, we also propose an adaptive ghost convolution (AGC) to replace part of the ordinary convolution. Using the PReLu activation function instead of the ReLu activation function in the ghost module, AGC can effectively improve the overfitting problem of the model and the quality of the repaired image without increasing the computational cost. The proposed model has undergone extensive experiments on several public datasets, and the results show that our method is superior to the state-of-the-art methods.

Keywords

Self-Attention, Adaptive Ghost Convolution, Image Inpainting, Large Missing Area, Artificial Intelligence, Deep Learning

1. Introduction

Image inpainting aims to use the known information about an image to reconstruct the missing or damaged part of the image to generate a new image. It is difficult to judge whether the new image. It is widely used to repair damaged images, remove certain objects from the image, and remove watermarks or subtitles from the image.

With the development of artificial intelligence and deep learning, excellent results have been achieved in the field of computer vision. Related convolutional neural networks (CNN) and generative adversarial

*Corresponding Author: Shiming He (smhe_cs@csust.edu.cn)
[1]School of Computer and Communication Engineering, Changsha University of Science & Technology, Changsha, China
[2]Department of Computer Science, Community College, King Saud University, Riyadh, Saudi Arabia
[3]Department of Biomedical Engineering, University of Reading, Reading, UK

network (GAN) have also been applied to the field of image restoration, and they have achieved remarkable results. Currently, most image inpainting methods are aimed at repairing damaged images of regular missing areas and irregular small missing area. Pathak et al. [1] proposed a context encoder, which first used a deep generative network to deal with an image inpainting model. It used the predictive ability of the autoencoder to predict the missing content of the damaged image based on the known image information around the damaged image. Iizuka et al. [2] added the improvement of global and local discriminators on the basis of the context encoder, which repaired images with any shape mask. Based on the context encoder, Yang et al. [3] proposed a multi-scale neural patch synthesis method according to the joint optimization of image content and texture constraints. These image inpainting methods are reasonable to use for damaged images of small missing area, because the pixels in a local area are strongly correlated and can be inferred from their surroundings. As the missing area of damaged images becomes larger, and the distance between the known pixel and the unknown pixel increases, however, these correlations weaken, and the constraint on the hole center eases. Therefore, the information in the known area is not helpful for the inpainting of the hole center pixel, generating semantically ambiguous inpainting results.

As such, many studies on image inpainting methods began to research on the damaged images of large missing area. Liu et al. [4] used a partial convolution strategy to repair the image and proposed a large irregular mask dataset. This method involves performing image restoration on irregularly shaped holes first. Image inpainting methods [5–10] are all two-stage network structures: predicting the structure information, and then performing content reconstruction. However, the error caused by the first-stage structure prediction can easily have an adverse impact on the second-stage content prediction, resulting in a poor final repair effect. Therefore, methods [11–16] use an end-to-end GAN structure. Nonetheless, the model based on the GAN structure is prone to overfitting problems during training, and it is difficult to conduct network training.

To overcome the limitations of the methods above, Li et al. [17] designed a recurrence feature reasoning network (RFR-Net) that used the shared feature inference module to repair progressively missing images with large mask areas and irregular shapes. However, it still produced unreasonable repair structures and some diamond texture. Therefore, in order to achieve better visual effects and reduce feature redundancy, we add self-attention mechanism and adaptive ghost convolution (AGC) to RFR-Net and propose recurrent feature reasoning based on self-attention for image inpainting (SA-RFR). Self-attention enhances the CNN's ability to perceive image size. Then, PReLU can effectively improve the overfitting problem of the model without increasing the amount of calculation and perform network training better. Therefore, we use the PReLU activation function instead of ReLU activation function in the ghost module (GM) [18] and design AGC that enhances the obtained feature maps. It can use fewer parameters to generate the same number of feature maps as the ordinary convolutional layer, and then become integrated into other networks.

Our contributions can be summarized as follows:

1. We introduce a self-attention mechanism to enhance the correlation between known pixels and unknown pixels and strengthen useful local texture features and similar block features. The repair effect of damaged images in large missing areas is improved, and the details of the repaired content are clearer.
2. We propose AGC instead of ordinary convolution. AGC enhances the obtained feature map and reduces feature redundancy. It can improve the quality of image restoration while reducing computational cost.
3. We analyze our model in terms of efficiency and performance and show the advantages of our network over several latest methods in public datasets.

The rest of this paper is organized as follows: Section 2 provides a brief review of related inpainting methods; Section 3 describes the proposed approach and loss functions in detail; Section 4 explains the experiments conducted for this work, the experimental comparison with other state-of-the-art methods, and the model analysis; and Section 5 summarizes this study.

2. Related Work

In recent years, a large number of deep learning-based algorithms in image inpainting have been proposed, achieving excellent results. Pathak et al. proposed a context encoder [1], i.e., an encoder-decoder architecture, as the earliest image inpainting method based on deep learning. Iizuka et al. [2] added the improvement of global and local discriminators on the basis of the context encoder, which repaired any shape mask. Yang et al. [3] proposed a multi-scale neural patch synthesis method based on the joint optimization of image content and texture constraints on the basis of the context encoder. It not only preserved the context structure but also produced high-frequency details. Yu et al. [5] proposed a contextual attention layer to extract features that approximated the area to be repaired from a distant area. It not only synthesized novel image structures but also explicitly used surrounding image features as references during network training. Yu et al. [6] introduced gated convolution to learn the dynamic feature selection mechanism of each channel at each spatial position, thereby improving the quality of inpainting images with arbitrary shapes of masks. The first step was to predict the foreground contour, and the second step was to repair the missing area content based on the foreground contour. It solved challenging inpainting scenarios involving foreground and background pixel prediction. Because the images in the local area have strong correlation, the pixels can be inferred from the surrounding environment. These methods are reasonable to use for repairing small or narrow defects. As the damaged area of the image becomes larger and the distance between the known pixel and the unknown pixel increases, however, these methods produce blurry and unreasonable repair results.

Therefore, many novel methods have emerged in recent years. Liu et al. [4] proposed a large irregular mask dataset as well as partial convolutions to repair any non-central and irregular damaged areas. Nazeri et al. [8] suggested a two-stage adversarial model edge connection network (EdgeConnect) consisting of an edge generator and an image restoration network. The edge generator generated predicted edges in the missing areas of the image, and the image completion network used the predicted edges as a priori to fill the missing areas. EdgeConnect can better repair the details of the filled area. Liu et al. [9] proposed a deep generative model method with context semantic attention (CSA), which performed more efficient inpainting by processing the semantic correlation between the void features. Qin et al. [12] suggested a novel multi-scale attention network (MSA-Net) to fill the irregular missing regions where a multi-scale attention group (MSAG) with several multi-scale attention units (MSAUs) is introduced for fully analyzing the features from shallow details to high-level semantics. Gupta and Kishore [13] considered and audited numerous distinct algorithms available for image inpainting and clarified their methodology. Yang et al. [15] developed a multi-task learning framework that attempted to combine image structure knowledge to assist in image inpainting. They proposed a novel pyramid structure loss to supervise structure learning and embedding. Wang et al. [16] suggested a new image inpainting method for large irregular masks that introduced a multi-stage attention module and then used a partial convolution strategy to repair the image in a rough to fine way. Li et al. [17] designed an RFR-Net composed mainly of a plug-and-play recursive feature reasoning module and a knowledge consistent attention (KCA) module that can effectively repair damaged images missing in a large area.

3. Methodology

In order to repair the damaged images of a large missing area, we propose SA-RFR, which conducts training in an end-to-end manner. There are ground truth I_{gt} and binary mask M whose value of known pixels is 0 and value of unknown pixels is 1. In this way, damaged image I_{in} is obtained from the ground truth as $I_{in} = I_{gt} \odot (1 - M)$. SA-RFR takes $[I_{in}, M]$ as input, and the predicted image is I_{out}. In the next subsection, we introduce the network architecture and three modules of SA-RFR in detail.

3.1 Network Architecture

As shown in Fig. 1, SA-RFR consists of three modules: the pretreatment module, the feature reasoning module, and the adaptive ghost fusion module. The damaged image and its mask are fed into the SA-RFR. The feature map is extracted, and the mask area is judged and updated by the pretreatment module. The feature reasoning module then infers the content of the damaged image in the mask area to generate a repaired feature map. The pretreatment module and the feature reasoning module alternately recur six times, and the repaired feature map is saved in each recurrence. Finally, in the adaptive ghost fusion module, all repaired feature maps get merged into a fixed feature map and generate a predicted image. The details of the three modules are as follows.

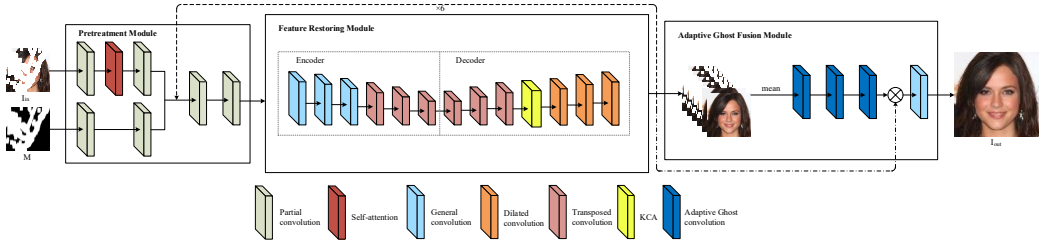


Fig. 1. Framework of our method.

3.1.1 Pretreatment module

The pretreatment module performs four partial convolution operations on the input image and the mask. The partial convolution aims to identify the area to be updated in each recurrence. For inpainting large continuous holes of damaged images, however, the known information of the image is scarce because the mask area is large; hence the lack of constraints for the hole center. In order to strengthen the long-distance correlation between the known and unknown pixels of the input image, a self-attention mechanism layer is added after the first partial convolution. As a supplement to the convolution, the self-attention module helps with modeling long-range, multi-level dependencies via image regions. It can correlate local details and relatively distant details, enriching the content information in the inpainting area. The mask and feature map are updated and saved in each recurrence, whose process is as follows:

$$X_1, M_1 = W_{p_1}(I_{in}, M) \tag{1}$$

$$X_2 = SA(X_1) \tag{2}$$

$$X_3, M_2 = W_{p_2}(X_2, M_1) \tag{3}$$

$$X_4, M_3 = W_{p_4}\left(W_{p_3}(X_3, M_2)\right) \tag{4}$$

where W_{p_i} is the i-th partial convolution and SA is the self-attention mechanism operation, with X_i, M_i as the corresponding feature maps and masks outputted by the i-th operation.

Self-attention: In CNN, the size of each convolution kernel is very limited; therefore, each convolution covers only a small neighborhood around the pixel. For obtaining distant features, multi-layer convolution and pooling operations are exploited, making the height and width of the feature map smaller and smaller. Since multiple layers of mapping are required, the more layers there are, the larger the area covered by the convolution kernel to be mapped back to the original image; hence the difficulty in capturing the expected features. The self-attention mechanism can take advantage of distant area information. Each location can be combined with the information of related areas to ensure the regional consistency of the inpainting image. Therefore, we apply the self-attention mechanism to enhance the

correlation between known pixels and unknown pixels, so that the details of the inpainting content are clearer and the edges are smoother. The self-attention mechanism operates as shown in Fig. 2.

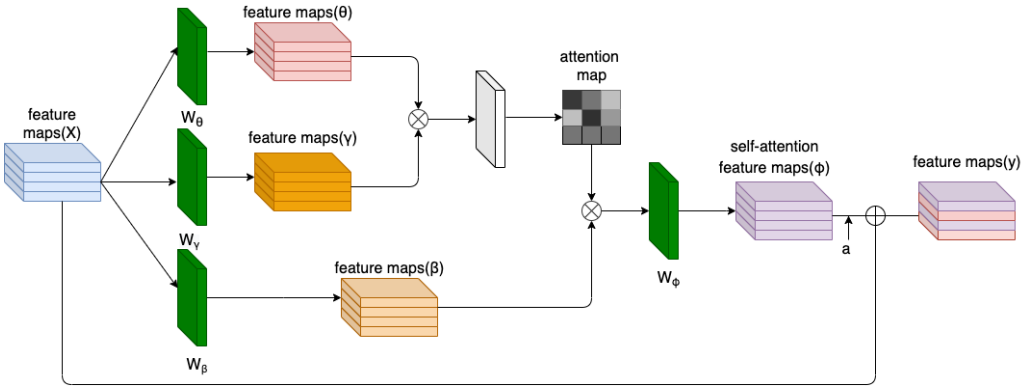


Fig. 2. Self-attention mechanism.

Suppose the feature map is X; two 1×1 convolutions are used for linear transformation and channel compression, and the output feature maps are θ and γ.

$$\theta = W_\theta(X), \gamma = W_\gamma(X) \tag{5}$$

where W_θ and W_γ are the 1×1 convolution.

Then we reshape θ and γ into matrix form, transpose and multiply them, and pass them through the softmax activation function to derive the attention map.

$$S_{j,i} = \sigma\left(\theta(X_i)^T \gamma(X_j)\right) \tag{6}$$

where $S_{j,i}$ is the model's attention to the position of i-th when synthesizing the j-th area and σ is the softmax operation.

Then X is linearly transformed through a 1×1 convolution, and the number of channels remains the same. Afterward, it is multiplied and added with the attention map before a 1×1 convolution to obtain the self-attention feature maps.

$$\beta = W_\beta(X) \tag{7}$$

$$g_{i,j} = S_j{}^T \beta(X_i) \tag{8}$$

$$\phi = W_\phi(g_{i,j}) \tag{9}$$

where β is the output feature map, W_β is a 1×1 convolution operation, the i-th row of $\beta(X_i)$ is all pixel values of the i-th channel, S is the attention map, S_j is the j-th column of the attention map representing the influence of all pixels on the j-th pixel, $g_{i,j}$ in row i and column j is the pixel value of the j-th pixel of the i-th channel of the feature map weighted by the attention map, ϕ is the self-attention feature map, and W_ϕ is a 1×1 convolution operation

Finally, the self-attention feature map and the original feature map are weighted and summed as the final output.

$$y = a\phi + X \tag{10}$$

where a is a weight parameter updated by backpropagation.

3.1.2 Feature reasoning module

The pretreatment module judges the area of the damaged image to be repaired, and then the updated feature map is fed into the feature reasoning module that seeks to use the known information to repair the feature map with high-level semantic features and to generate an inpainting result with reasonable structure and rich texture. The feature reasoning module has an encoder-decoder structure: the encoder includes six down sampling, whereas the decoder includes three up sampling, a KCA module, and three transposed convolutions. Unlike ordinary attention, which is calculated independently, KCA is the weighted sum of the scores obtained from six recurrences. The feature reasoning module, which takes X_4 and M_4 as the input, is expressed as follows:

$$Encoder \begin{cases} X_5 = W_3\left(W_2(W_1(X_4))\right) \\ X_6 = W_{d_3}\left(W_{d_2}\left(W_{d_1}(X_5)\right)\right) \end{cases} \tag{11}$$

$$Decoder \begin{cases} X_7 = W_{d_6}\left(W_{d_5}\left(W_{d_4}(X_6)\right)\right) \\ X_8 = KCA(X_7) \\ F_i = W_{T_3}\left(W_{T_2}\left(W_{T_1}(X_8)\right)\right) \end{cases} \tag{12}$$

where W_i is the i-th convolution, W_{d_i} is the i-th dilated convolution, W_{T_3} is the i-th transposed convolution, and F_i is the feature map generated in the i-th recurrence saved in each recurrence.

3.1.3 Adaptive ghost fusion module

The pretreatment module and the feature reasoning module alternately recur six times until the mask area is completely filled. Then the adaptive ghost fusion module first merges the feature maps saved by the six recurrences. Because the mask regions of different feature maps saved are not the same, merging feature maps can effectively avoid too abrupt values in certain positions that result in inconsistencies in the texture or structure of the predicted image. We use AGC instead of convolution as shown in Fig. 3, and this can reduce feature redundancy, deepen the network, and achieve better inpainting effect. Through three AGC layers, the merged feature map and the feature map generated by the sixth recurrence are concatenated together, and the predicted image is then outputted through a ReLU layer. Let the feature map saved in the i-th recurrence be F_i, and F is the merged feature maps.

$$F = \frac{1}{6}\sum_{i=1}^{6} F_i \tag{13}$$

The operations of the adaptive ghost fusion module are as follows:

$$y_1 = R\left(B\left(C_{Ag_1}(F)\right)\right) \tag{14}$$

$$y_2 = R\left(B\left(C_{Ag_2}(y_1)\right)\right) \tag{15}$$

$$y_3 = B\left(C_{Ag_3}(y_2)\right) \tag{16}$$

$$Y = R(F + y_3) \tag{17}$$

where C_{Ag} is the AGC layer, B is the batch normalization (BN) layer, R is the ReLU activation function, y_i is the result of the i-th layer, and Y is the output of the adaptive ghost fusion module.

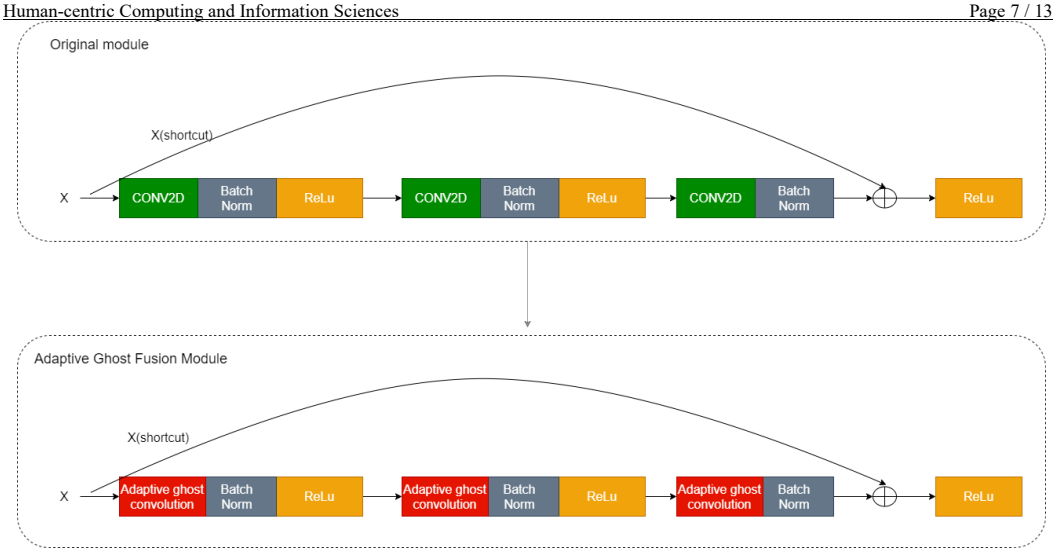


Fig. 3. Changed adaptive ghost fusion module.

Adaptive ghost convolution: Because the feature map generated by convolution contains a lot of redundant feature information, many feature maps are similar, resulting in the generation of redundancy and unnecessary information; the quality of the inpainting results is also poor. Furthermore, most of the existing methods are a two-stage network structure or a GAN-based network structure. This type of network structure consumes computational resources, and the network is not well-trained and is prone to overfitting. Therefore, we designed an AGC as shown in Fig. 4. AGC uses the PReLU activation function instead of the ReLU activation function in the ghost module. PReLU can effectively improve the overfitting problem of the model, and it also offers faster convergence and better network training.

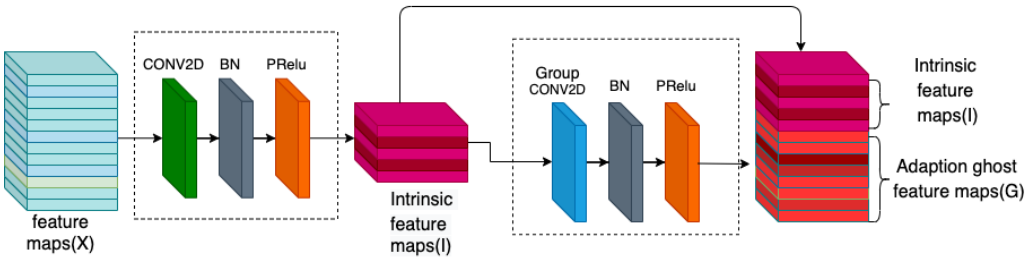


Fig. 4. Adaptive ghost convolution.

AGC is similar to the ghost module, which is divided into three steps: conventional convolution, ghost feature maps generation, and feature maps stitching.

First, feature map Z is taken as input, with intrinsic feature maps I obtained through conventional convolution W.

$$I = W(X) = P\left(B\left(C(Z)\right)\right) \tag{18}$$

where C is the conventional convolution layer, B is the BN layer, and P is the PReLU activation function.

Adaptive ghost feature maps G_{ij} are then generated by group convolution W_G.

$$G_{ij} = W_{G_{j,i}}(I_i) = P\left(B\left(C_G(X)\right)\right), i \in 1,2,\cdots,m, j \in 1,2,\cdots,s \tag{19}$$

where $P\left(B\left(C_G(X)\right)\right)$ is the process of group convolution W_G, C_G is the group convolution layer, I_i is the i-th feature map in the intrinsic feature maps, i is the sequence number in the m intrinsic feature maps, and j is the j-th linear transformation of the feature map in each intrinsic feature map, and it is convolved one feature map at a time.

Finally, the intrinsic feature map I obtained in the first step and the adaptive ghost feature map G obtained in the second step are concatenated together to obtain the final feature map Y.

$$Y = cat(I, G) \tag{20}$$

3.2 Loss Functions

We use a hybrid loss similar to that in [17] to repair the image, which includes four loss functions: perceptual loss, style loss, L_1 as loss of unknown regions, and L_1 as loss of known regions.

Perceptual loss: It compares the feature map generated from the ground truth and the feature map generated from the predicted image and makes the texture content and the global structure and other high-level information closer.

$$L_{perceptual} = \sum_i \left|\left|\phi_i(I_{gt}) - \phi_i(I_{out})\right|\right|_1 \tag{21}$$

where ϕ_i is the i-th feature map generated in VGG16.

Style loss: The purpose is to maintain the color and pattern consistency between the predicted image and the ground truth. First, the Gram matrix is calculated.

$$G_j^{\phi}(x)_{c,c'} = \frac{1}{C_j H_j W_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} \phi_j(x)_{h,w,c} \phi_j(x)_{h,w,c'} \tag{22}$$

where $G_j^{\phi}(x)_{c,c'}$ is the inner product of the feature map of each channel c and the feature map of each channel c', j is the j-th layer of the network, the input feature map is defined as $\phi_j(x)_{h,w,c}$, and the size is $C_j H_j W_j$.

Therefore, style loss L_{style} can be defined as:

$$L_{style}^{\phi,j}(I_{gt}, I_{out}) = \left|\left|G_j^{\phi}(I_{gt}) - G_j^{\phi}(I_{out})\right|\right|_F^2 \tag{23}$$

Hole loss: L_{hole} calculates L_1 as loss of the unknown area.

$$L_{hole}(I_{gt}, I_{out}) = \frac{1}{n} \sum_i^n M_i * \left|I_{gt_i} - I_{out_i}\right| \tag{24}$$

Valid loss: L_{hole} calculates L_1 as loss of the known area.

$$L_{valid}(I_{gt}, I_{out}) = \frac{1}{n} \sum_i^n (1 - M_i) * \left|I_{gt_i} - I_{out_i}\right| \tag{25}$$

In summary, our total loss function is:

$$Loss = \lambda L_{perceptual} + \mu L_{style} + \eta L_{hole} + \gamma L_{valid} \tag{26}$$

The combination of loss functions in our model requires fewer parameters to be updated, and it can also achieve efficient training.

4. Experimental Results and Analysis

4.1 Experimental Settings

4.1.1 Training setup and strategy

We evaluate SA-RFR on two public datasets: Places2 and CelebA. We randomly partition Places2 into 36k images for training and 100 images for testing, and CelebA into 29k images for training and 100 images for testing. Images in Places2 and CelebA are resized to 512×512 and 256×256, respectively, during training and testing. We use the irregular mask dataset provided by [6]. For our experiment, we empirically set λ=0.1, μ=180, η=6, and γ=1 in Equation (26). The training procedure involves using Adam optimizer. We divide the training model into two parts: normal training and fine-tuning training. We set the learning rate to $2e^{-4}$ for normal training and to $5e^{-5}$ for fine-tuning training. The batch size is 2. We apply PyTorch framework to implement our model and train it using NVIDIA GeForce RTX 3090 GPU (24 GB memory). The operating system is Ubuntu16.08, the CPU is Intel i5-10400F, and the memory size is 32 GB.

We compare our approach with several state-of-the-art methods including EdgeConnect [8], DF-Net [19], PIC-Net [20], and RFR-Net [17]. We conduct qualitative analysis and quantitative analysis to demonstrate the superiority of our method. Finally, we perform ablation experiments on the CelebA dataset to check the design details of SA-RFR.

4.1.2 Evaluation measures

We measure the quality of our results using the following metrics: peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM).

PSNR: PSNR is an index that is widely used to evaluate the distortion of reconstructed images objectively. As shown in Equation (27), PSNR calculates the similarity based on the mean square error (MSE) between the repaired image and the original image, and the unit is decibel (dB). The greater the PSNR value is, the less the distortion of the repaired image and the better the effect.

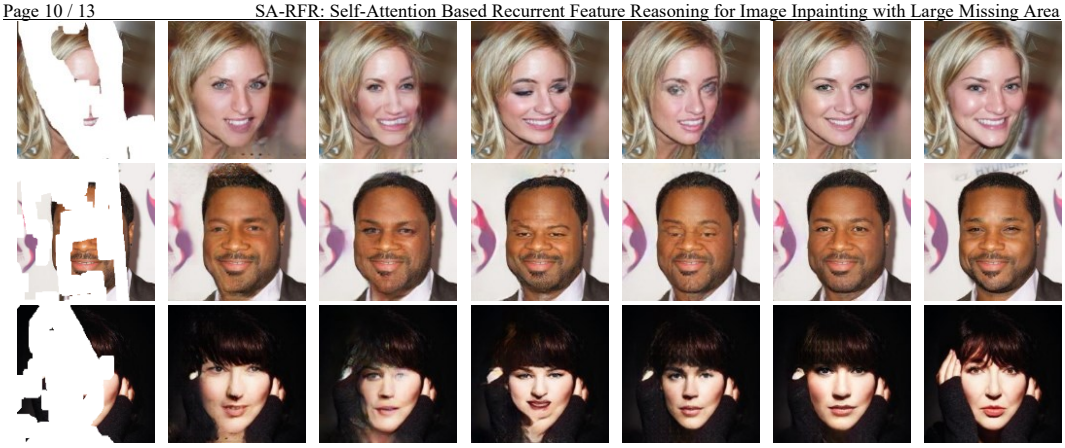
$$PSNR = 10 \cdot log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{27}$$

SSIM: Objects in natural scenes have strong structural features, and such structural feature and the illumination information are independent of each other. As shown in Equation (28), SSIM estimates the similarity between the original image and the repaired image by combining the structure information, luminance information, and contrast information of the image. The greater the SSIM value is, the better the effect.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{28}$$

4.2 Qualitative Comparisons

As shown in Figs. 5 and 6, our inpainting results are significantly better than the state-of-the-art methods especially for large continuous holes of damaged images. For the CelebA dataset, our results have more real details and fewer artifacts compared to the comparison method. In addition, we compare our method and the latest method on the Places2 dataset. Our repair of the image structure is better than the other methods, and the image generated has more reasonable and beautiful results in semantics thanks to the self-attention mechanism and the AGC.



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

Fig. 5. Results on CelebA: (a) masked input, (b) PIC-Net, (c) EdgeConnect, (d) DF-Net, (e) RFR-Net, (f) SA-RFR, and (g) ground truth.

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

Fig. 6. Results on Places2: (a) masked input, (b) PIC-Net, (c) EdgeConnect, (d) DF-Net, (e) SA-RFR, and (f) ground truth.

4.3 Qualitative Comparisons

We measure the quality of our results using the PSNR and SSIM; the higher the SSIM and PSNR values are, the better the effect. Table 1 lists the results with different ratios of irregular masks for the two datasets. As shown in Table 1, our method produces excellent results on the Places2 and CelebA datasets. When the mask ratio is 10%–20%, the gain of PSNR and SSIM values is small; when the mask ratio is >50%, however, PSNR increases by 1.1. This is because self-attention enhances the correlation between the known and missing regions of the image, and AGC improves the features. Therefore, our PSNR and SSIM are enhanced.

Table 1. Quantitative results on two testing datasets

		Places2			CelebA		
		10%–20%	**30%–40%**	**>50%**	**10%–20%**	**30%–40%**	**>50%**
PSNR	PIC-Net	27.14	21.72	17.17	30.67	24.74	20.34
	EdgeConnect	26.41	23.11	18.35	33.04	25.72	21.24
	DF-Net	27.93	23.55	19.04	33.70	26.69	21.38
	RFR-Net	27.75	22.63	18.92	33.56	27.15	22.35
	SA-RFR (ours)	**28.09**	**23.63**	**20.12**	**34.24**	**27.69**	**23.26**
SSIM	PIC-Net	0.932	0.786	0.494	0.965	0.881	0.672
	EdgeConnect	0.913	0.806	0.553	0.957	0.856	0.728
	DF-Net	0.926	0.821	0.682	0.966	0.872	0.740
	RFR-Net	**0.939**	0.819	0.596	0.966	0.877	0.750
	SA-RFR (ours)	0.937	**0.831**	**0.690**	**0.968**	**0.889**	**0.785**

The best values are marked in bold.

4.4 Ablation Study

In order to synthesize the effect of self-attention and AGC module in the SA-RFR, we conduct an ablation study on the CelebA dataset. The comparison results are shown in Table 2.

Table 2. Influence of adding self-attention and AGC on the network inpainting effect

Model	SA	AGC	GM	CelebA (mask >50%)	
				PSNR	SSIM
RFR-Net				22.35	0.750
SA-RFR(-AGC)	✓			22.88	0.766
SA-RFR(-SA)		✓		22.92	0.772
SA-RFR(-AGC+GM)	✓		✓	22.71	0.766
SA-RFR	✓	✓		23.26	0.785

Effectiveness of self-attention: To investigate the effect of self-attention, we train SA-RFR without self-attention (SA-RFR(-SA)). As shown in Fig. 7(d), SA-RFR(-SA) produces unnatural textures. In contrast, SA-RFR produces the repaired content with clearer details and smoother edges. Furthermore, we compare the quantitative performance of these two models with mask ratio of >50% as shown in Table 2. Self-attention can use the information of a distant area to ensure the regional consistency of the predicted image. Therefore, SA-RFR can generate better texture details. SA- RFR is better than SA-RFR(-SA) in the qualitative or quantitative aspects.

(a) (b) (c) (d) (e)

Fig. 7. Comparison results of different modules: (a) masked input, (b) RFR-Net, (c) SA-RFR(-AGC), (d) SA-RFR(-SA), and (e) ground truth.

Effectiveness of AGC: In order to prove the effectiveness of AGC, we train SA-RFR without adaptive ghost convolution (SA-RFR(-AGC)) for comparison. As shown in Fig. 7, SA-RFR(-AGC) produces results with unreasonable structure, especially the parts of the eyes and eyebrows. However, SA-RFR has better facial features. At the same time, our model with AGC is better than RFR-Net without this module in the numerical indicators of PSNR and SSIM, as shown in Table 2.

Furthermore, our AGC is improved on the ghost module. In order to prove the effectiveness of the improvement, we replace AGC with GM in SA-RFR (SA-RFR(-AGC+GM)). As shown in Fig. 8(b), the lip part of the image in the first row has obvious artifacts and unreasonable structures. The image generated in the second row has obvious fish scale artifacts around the neck. In contrast, the images generated by AGC (Fig. 8(c)) are more realistic and reasonable. In the last two rows in Table 2, under the premise of the same use of self-attention, our proposed AGC is significantly better than the value of the ghost module; thus proving the effectiveness of AGC.

(a) (b) (c) (d)

Fig. 8. Comparison results of AGC and ghost module: (a) masked input, (b) SA-RFR(-AGC+GM), (c) SA-RFR, and (d) ground truth.

5. Conclusion

We have proposed SA-RFR for the damaged images missing in a large area. SA-RFR uses self-attention and AGC, which enhance the correlation between known pixels and unknown pixels and improve the quality of the repaired image without increasing computational cost. Experiments have verified the effectiveness of our proposed method.

Author's Contributions

Conceptualization, SH, JW. Investigation and methodology, LW, OA, AT. Project administration, SH. Supervision, JW. Writing of the original draft, LW. Writing of the review and editing, SH, AT, RSS. Software, LW. Validation, SH, RSS. Formal analysis, JW, OA. Data curation LW, AT.

Funding

This work was supported in part by the National Natural Science Foundation of China (No. 61802030), King Saud University (Riyadh, Saudi Arabia) through the Researchers Supporting Project (No. RSP-2021/102), Natural Science Foundation of Hunan Province (No. 2020JJ2029, 2020JJ5602), Research Foundation of Education Bureau of Hunan Province (No. 19B005), and International Cooperative Project for "Double First-Class" CSUST (No. 2018IC24).

Competing Interests

The authors declare that they have no competing interests.

References

[1] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context Encoders: Feature Learning by Inpainting," in *Prodeedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 2536-2544.

[2] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics*, vol. 36, no. 4, article no. 107, 2017. https://doi.org/10.1145/3072959.3073659

[3] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, “High-resolution image inpainting using multi-scale neural patch synthesis,” in *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 4076-4084.

[4] 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 (ECCV)*, Munich, Germany, 2018, pp. 85-100.

[5] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 5505-5514.

[6] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," in *Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV*), Seoul, Korea, 2019, pp. 4470-4479.

[7] D. Cao, X. Ren, M. Zhu, and W. Song, “Visual question answering research on multi-layer attention mechanism based on image target features,” *Human-centric Computing and Information Sciences*, vol. 11, article no. 11, 2021. https://doi.org/10.22967/HCIS.2021.11.011

[8] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "EdgeConnect: structure guided image inpainting using edge prediction," in *Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Seoul, Korea, 2019, pp. 3265-3274.

[9] H. Liu, B. Jiang, Y. Xiao, and C. Yang, "Coherent semantic attention for image inpainting," in *Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV),* Seoul, Korea, 2019, pp. 4169-4178.

[10] X. Zhu, K. Guo, H. Fang, L. Chen, S. Ren, and B. Hu, "Cross view capture for stereo image super-resolution," *IEEE Transactions on Multimedia*, 2021. https://doi.org/10.1109/TMM.2021.3092571

[11] X. Zhu, K. Guo, S. Ren, B. Hu, M. Hu, and H. Fang, "Lightweight image super-resolution with expectation-maximization attention mechanism," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1273-1284, 2022.

[12] J. Qin, H. Bai, and Y. Zhao, "Multi-scale attention network for image inpainting," *Computer Vision and Image Understanding*, vol. 204, article no. 103155, 2021. https://doi.org/10.1016/j.cviu.2020.103155

[13] M. Gupta and R. R. Kishore, "Different techniques of image inpainting," in *Computational Methods and Data Engineering*. Singapore: Springer, 2021, pp. 93-104.

[14] H. H. Bu, N. C. Kim, and S. H. Kim, “Content-based image retrieval using a combination of texture and color features,” *Human-centric Computing and Information Sciences*, vol. 11, article no. 23, 2021. https://doi.org/10.22967/HCIS.2021.11.023

[15] J. Yang, Z. Qi, and Y. Shi, “Learning to incorporate structure knowledge for image inpainting,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, New York, NY, 2021, pp. 12605-12612.

[16] N. Wang, S. Ma, J. Li, Y. Zhang, and L. Zhang, “Multistage attention network for image inpainting,” *Pattern Recognition*, vol. 106, article no. 107448, 2020. https://doi.org/10.1016/j.patcog.2020.107448

[17] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, “Recurrent feature reasoning for image inpainting,” in *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* Seattle, WA, 2020, pp. 7757-7765.

[18] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, “GhostNet: more features from cheap operations,” in *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* Seattle, WA, 2020, pp. 1577-1586.

[19] X. Hong, P. Xiong, R. Ji, and H. Fan, “Deep fusion network for image completion,” in *Proceedings of the 27th ACM International Conference on Multimedia*, Nice, France, 2019, pp. 2033-2042.

[20] C. Zheng, T. J. Cham, and J. Cai, “Pluralistic image completion,” in *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, 2019, pp. 1438-1447.