Srgan-Sa: Super-Resolution Generative Adversarial Network with Integrated Attention Mechanisms

Kangdi Huang[1], Yingce Xu[2], Zhe Zhang[3]*

[1]Faculty of Printing, Packing and Digtial Media Technology, Xi'An University of Technology, Shaanxi, China
[2]International Education School, Hebei University of Economics and Business, Shijiazhuang, Hebei, China
[3]School of International Education, Guangxi University of Science and Technology, Liuzhou, Guangxi, China

Abstract. With the development of the times, this paper proposes a dynamic spatial attention enhancement SRGAN-SA model to better solve the above problems, such as the loss of high-frequency details in traditional image super-resolution reconstruction (SISR) methods, and the insufficient attention of SRGAN global feature processing mechanism to key areas. Based on the deep learning module, the model designs a learnable 7 * 7 convolution to generate dynamic attention maps. It's embedded in the residual module of SRResNet generator to realize adaptive focus and feature reweighting of regions. The model adopts the confrontation training paradigm, combines VGG perception loss and confrontation loss optimization generator, and uses depth convolution discriminator to enhance texture identification. Experimental results show that compared with SRGAN, the PSNR of this method on Set5 test set is increased by 1.747dB, reaching 31.6381, and the SSIM is increased by 0.0377, reaching 0.8863. The visualization results show that the improved model can effectively suppress artifacts. The ablation experiment further verified that the contribution rate of the dual pool strategy and the dynamic convolution design to the performance improvement reached 0.82 dB and 0.65 dB respectively. It provides a new solution for the deep integration of attention mechanism and confrontation training.

1 Introduction

With the rapid advancement of digital media technologies, single image super-resolution (SISR) reconstruction has demonstrated significant value in remote sensing imaging, medical image enhancement, and video restoration. Early traditional methods, such as bicubic interpolation [1] and sparse coding-based optimization algorithms [2], relied on handcrafted features to map low-resolution (LR) images to high-resolution (HR) counterparts. However, limited by their model expressiveness, these approaches struggled to recover high-frequency details, often producing blurred outputs or artifacts. The breakthrough in deep learning led

* 3220342091@stu.xaut.edu.cn

Dong et al. to propose the Super-Resolution Convolutional Neural Network (SRCNN), which first adopted an end-to-end convolutional neural network (CNN) to directly learn LR-to-HR mappings, significantly improving reconstruction accuracy [3]. Despite this progress, methods relying on mean squared error (MSE) loss functions tended to generate overly smooth images lacking realistic texture details. Subsequent improvements, such as the Very Deep Super-Resolution (VDSR) network [4], enhanced performance by increasing network depth, yet perceptual quality remained a bottleneck.

To address these limitations, Ledig et al. introduced the Super-Resolution Generative Adversarial Network (SRGAN), pioneering the integration of generative adversarial networks (GANs) into SISR [5]. Its generator, based on the residual structure of SRResNet, employed adversarial training to produce realistic details, while the discriminator distinguished generated images from real HR images. SRGAN combined VGG-based perceptual loss to optimize texture fidelity. However, its global feature processing mechanism inadequately focused on high-frequency information in critical regions, leading to local distortions in complex scenes. Enhanced models like ESRGAN removed batch normalization layers, introduced relativistic discriminators, and refined residual structures to improve detail representation, yet challenges persisted in multi-scale feature fusion and dynamic attention allocation [6].

Attention mechanisms, which dynamically allocate feature weights to enhance the modeling of key regions, have opened new avenues for SISR. For instance, Zhang et al. proposed the Residual Channel Attention Network (RCAN), leveraging channel attention modules to adaptively amplify high-frequency features [7]. Dai et al. designed the Second-Order Attention Network (SAN), optimizing attention weights through higher-order statistics [8]. Recent studies have integrated attention mechanisms with GAN frameworks. Wang et al. developed Edge-Aware SRGAN (EA-SRGAN) to enhance texture generation via spatial attention modules [9], while Chen et al. designed a multi-scale attention generator for dynamic weight allocation [10].

Existing spatial attention modules face several limitations: they predominantly rely on global average pooling, which inadequately integrates multi-dimensional contextual information such as high-frequency features captured by global max pooling. Additionally, the attention weight generation process often employs fixed-scale convolutions, limiting adaptability to multi-scale textures and constraining the reconstruction of intricate local details. Furthermore, current methods exhibit suboptimal integration of attention mechanisms with GAN-based adversarial training frameworks, resulting in restricted perceptual optimization capabilities for generators. These challenges collectively hinder the effective alignment of texture synthesis with high-frequency feature enhancement in complex super-resolution scenarios.

To address these challenges, our approach integrates multi-dimensional feature fusion by concatenating dual-path features derived from global max pooling and average pooling, enabling the capture of hierarchical spatial context across different levels. Furthermore, a learnable 7×7 convolutional layer is introduced to dynamically generate adaptive attention maps, which prioritize high-frequency regions for enhanced detail reconstruction. The optimized attention module is embedded into the residual blocks of SRResNet, synergizing VGG perceptual loss with adversarial training to refine the generator's perceptual quality, while a deep convolutional discriminator strengthens texture discrimination through adversarial optimization. Experimental results demonstrate significant improvements, with the method achieving gains of 1.747 dB in PSNR and 0.0377 in SSIM over SRGAN and EA-SRGAN on the Set5 dataset, underscoring the effectiveness of dynamic attention mechanisms in adversarial training frameworks for high-fidelity super-resolution tasks.

2 Method

The Generative Adversarial Networks have played a significant role in many fields, and their derivative model, SRGAN, has achieved remarkable results in SISR. The result of traditional methods (such as Bilinear interpolation or Bicubic interpolation) has the defect that the pixels that are far apart cannot establish connections. In contrast, the result of SRGAN displays more details of images. For the sake of enhancing the generation result of SRGAN, this paper add spatial attention (SA) based on SRGAN. This model is able to lead to better results by enabling important features to be used based on weights assigned.

2.1 Overall Structure

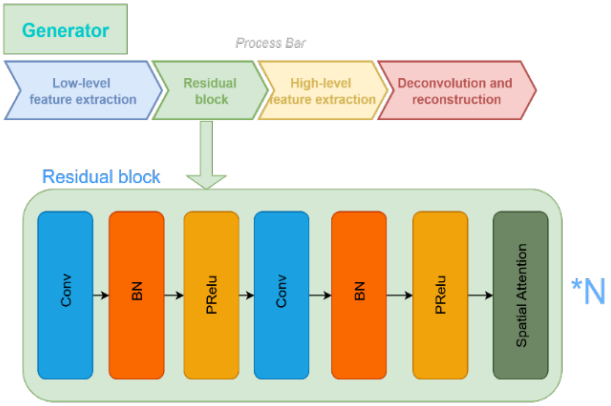

Figure 1 Residual block of Generator (Picture credit: Original)

As shown in Figure 1, the generator consists of a low-level feature extraction, residual blocks, high-level feature extraction and deconvolution and reconstruction. The innovation is that adds SA for residual blocks.

2.2 SRGAN-SA

This paper adopts SRResNet, which consists of several residual blocks (16 in this paper) and an upsampling module.

The input image first passes through a low-frequency information extraction layer (conv1), which uses a 9*9 convolutional kernel with 64 output channels and uses the PRelu activation function. After that, the image enters into the residual module, and each residual module consists of two convolutional layers plus batch normalization and PRelu activation function, after which the high-frequency information can be stripped out of the image. This layer can strip out the high frequency information of the image, and at the same time, Skip Connection is used to retain the low frequency information. In this paper, spatial attention is newly added to the residual module in order to separate the more contributing feature blocks. The image enters the next layer for the fusion of the high-frequency information, which is carried out by a convolutional layer (conv2) and batch normalization is used; after that, the image enters the next layer to start the image restoration task, first upsampling is performed to increase the resolution of the input image to that of the target image, through a convolution

operation and Pixel Shuffle, and after PRelu activation is output to the reconstruction layer. The reconstruction layer restores the number of channels of the image to 3 through a convolution operation (conv3) using a convolutional kernel of 9*9.

The discriminator adopts a structure like VGG, used for image discrimination. It comprises multiple convolutional layers, each followed by batch normalization and a LeakyReLU activation function. The channels gradually increased from 64 to 512. During this process of channel expansion, there are convolutional layers with a stride of 2 for downsampling. The last layer is the classifier, which flattens and classifies by Linear.

2.3 Spatial Attention

Spatial Attention first does a global max-pooling and global average mooling based on channel to obtain two feature maps whose size is H×W×1, and then these two feature maps are performed a splice operation based on channel. Then after a 7×7 convolution operation, the dimension is reduced to H×W×1.In the last, the spatial attention features are generated by sigmoid [11]. The specific processes are shown in Equation (1), Equation (2) and Equation (3).

$$F_{avg} = AvgPool(F) F_{max} = MaxPool(F) \tag{1}$$

$$M(F) = \sigma\left(f^{7\times7}\left(\left[F_{avg}; F_{max}\right]\right)\right) \tag{2}$$

$$Out = M(F) \odot F \tag{3}$$

$f^{7\times7}$ represents a convolution operation with a convolution kernel size of 7×7, σ denotes a sigmoid operation, $\odot$ represents an element-by-element multiplication, and after that, each element of the matrix is assigned an attention weight.

3 Experiment

The dataset adopted is a part of Imagenet. Before training, this paper crop the dataset into 128*128 size. This paper gets results by training 50 epochs on RTX 4090.

3.1 Loss Function

In SRGAN, a new loss function named perceptual loss function is defined which consists of content loss function and adversarial loss function according to their weights.

$$L_{content} = \frac{1}{n_l} \sqrt{\sum_{i,j} \left(\varphi_{i,j}^{(l)}(I) - \varphi_{i,j}^{(l)}(\hat{I})\right)^2} \tag{4}$$

The n_l in Equation(4) represents the number of pixel points corresponding to the feature maps in the Lth layer. $\varphi_{i,j}^{(l)}(I)$ and $\varphi_{i,j}^{(l)}(\hat{I})$ denote the feature maps obtained by the jth convolution of the reconstructed image and the original high-resolution image before the ith max pooling layer in the Lth layer, respectively.

$$L_G(I_{SR}) = -\ln\left(D(I_{SR})\right) \tag{5}$$

$$L_D(I_{SR}, I_{HR}) = -\ln\left(D(I_{HR})\right) - \ln\left(1 - D(I_{SR})\right) \tag{6}$$

$L_G(I_{SR})$ is the loss function of the generator obtained based on cross entropy, $L_D(I_{SR}, I_{HR})$ is the loss function of the discriminator, $D(I_{SR})$ denotes the probability that the image generated by the generator is an actual image.

$$L^{SR} = L_{content} + 10^{-3} L_G(I_{SR}) \tag{7}$$

L^{SR} is the perceptual loss function, $L_{content}$ is the content loss function, $L_G(I_{SR})$ is the adversarial loss function.

3.2 Experimental results and Analysis

Table 1 Test result

	PSNR	SSIM
SRGAN	29.8911	0.8486
SRGAN-SA	31.6381	0.8863

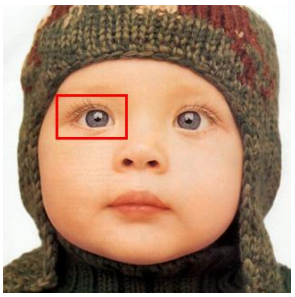

Figure 2 The Original Image (Picture credit: Original)

Figure 3 Output of SRGAN (Picture credit: Original)

Figure 4 Output of SRGAN-SA (Picture credit: Original)

From the results in Table 1, it can be seen that SRGAN-SA improves PNSR by 1.747 and SSIM by 0.0377 compared to SRGAN in Set5. Figures 2, 3, and 4 respectively represent the original image, the image generated by SRGAN, and the image generated by SRGAN-SA. Focusing on the eye region in Figures 2, 3, and 4, it can be seen that Figure 3 shows an eye-visible improvement in eye details compared to Figure 2.

4 Conclusion

The main findings of this study indicate that the introduction of a dynamic spatial attention mechanism into the SRGAN model has successfully enhanced the model's performance in image super-resolution reconstruction tasks. Experimental results demonstrate that the SRGAN-SA model achieves significant performance improvements over the original SRGAN on the Set5 test set. These improvements are not only reflected in objective metrics such as PSNR and SSIM, but also in subjective visual effects, showing better detail restoration capabilities and visual fidelity. This further proves that the dynamic spatial attention mechanism can effectively focus on key areas of the image, optimize high-frequency detail restoration, and improve the utilization of computational resources.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

1. R. Keys, "Cubic convolution interpolation for digital image processing," IEEE Trans. Acoust., Speech, Signal Process., 1981.
2. J. Yang, J. Wright, T. S. Huang and Y. Ma, "Image super-resolution via sparse representation," IEEE Trans. Image Process., 2010.
3. C. Dong, C. C. Loy, K. He and X. Tang, "Image super-resolution using deep convolutional networks," IEEE Trans. Pattern Anal. Mach. Intell., 2016.
4. J. Kim, J. K. Lee and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016.
5. C. Ledig, L. Theis, F. Huszár et al., "Photo-realistic single image super-resolution using a generative adversarial network," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017.
6. X. Wang, K. Yu, S. Wu et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," in Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops, 2018.
7. Y. Zhang, K. Li, K. Li et al., "Image super-resolution using very deep residual channel attention networks," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018.
8. T. Dai, J. Cai, Y. Zhang, S. Xia and L. Zhang, "Second-order attention network for single image super-resolution," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019.
9. H. Wang, Y. Li, Y. Wang and L. Liu, "EA-SRGAN: Edge-aware generative adversarial network for image super-resolution," IEEE Trans. Image Process., 2021.
10. Z. Chen, H. Zhang, Y. Wang and Q. Zhang, "Multi-scale attention network for image super-resolution," Neurocomputing, 2022.

11. C. Zhang, L. Zhu and L. Yu, "A review of attention mechanisms in convolutional neural networks," Comput. Eng. Appl., vol. 57, no. 20, pp. 64–72, 2021