FASRGAN: Feature Attention Super Resolution Generative Adversarial Network

Aditya Thaker, Akshath Mahajan, Adithya Sanyal, and Sudhir Bagul

Abstract The advent of multimedia-based communication requires high-speed, memory-efficient data storage of high quality images. One method for reducing the dimensions of an image while preserving the finer texture characteristics is image super resolution. This paper proposes a new model to deal with the reconstruction time and artifacts that GAN-based deep learning models face. To increase the speed, FASRGAN aims at reducing the complexity of generators. The network architecture builds on the ideas from ESRGAN and RAMS to derive a more efficient FASRGAN. This paper introduces the Residual Feature Attention Block (RRFAB) as a fundamental feature extraction block. Moreover, a relativistic discriminator is employed, inspired by relativistic GAN, that predicts the realness value of the generated image rather than a strict class value. Ultimately, training is performed on the model to improve on the content loss, which helps to converge the weights in the later training stages. The proposed model FASRGAN, achieves PSNR scores 31.5 and 31.04, SSIM scores 0.975 and 0.93 on Set5 and Set14 datasets, respectively, that are greater than the state of the art techniques and SSIM scores that are comparable to them.

Keywords Single image super resolution · Generative adversarial network · SRGAN · ESRGAN

1 Introduction

Single image super resolution is the task of re-creating high resolution images from their low resolution counterparts. Due to the large amount of information that humans perceive through visual media, super resolution has garnered a significant amount of interest in the field of computer vision.

The problem itself can be avoided by using cutting-edge equipment to capture visual material, however, this route is extremely costly and often infeasible. Gener-

A. Thaker · A. Mahajan · A. Sanyal (✉) · S. Bagul
Dwarkadas J. Sanghvi College of Engineering, Mumbai, Maharashtra, India
e-mail: adithyasanyal@gmail.com

S. Das et al. (eds.), *Advances in Data-Driven Computing and Intelligent Systems*,
Lecture Notes in Networks and Systems 891,
https://doi.org/10.1007/978-981-99-9524-0_18

ating and transmitting high quality images is highly resource intensive. This serves as a major disadvantage in many scenarios, such as cheap projects, quick implementation of prototypes, or other applications where generating high quality images may just be a small part of a larger process. To overcome these limitations, the process of upscaling low resolution images, known as image super resolution is proposed. These super resolution algorithms like nearest neighbor interpolation [1], bilinear interpolation [2], bicubic interpolation [3] principal objective was to approximate pixels in high resolution photographs utilizing information from adjacent pixels. However it was noted that the reconstructed images to use these methodologies lacked high-frequency details.

Numerous approaches to deep learning have sprouted in the past few years. Convolutional Neural Networks (CNN) is recommended to deal with super resolution, however it results in images having over smoothed details without sufficient focus on the finer details. Other widely recognized deep learning techniques for super resolution are entrenched in the concept of GANs. A Generative Adversarial Networks (GAN), is a deep structured learning paradigm wherein two neural networks vie with one another to enhance the precision of their forecasts.

Super resolution through the use of GANs has been pioneered by SRGAN [4], which trains the model through adversarial and content losses to provide visually pleasing images. Many GAN-based models that further improve SRGAN have been proposed in recent times, however these methods still have the presence of unwanted artifacts in their outputs and the metrics of these images can be improved.

In this paper, banking on the latest innovations in the field of super resolution, we propose a deep learning-based solution namely Feature Attention Super Resolution Generative Adversarial Networks (FASRGAN) which improves on ESRGAN. This system works superior to current approaches in both qualitative and quantitative assessments, done in terms of various metrics like Peak Signal-to-Noise Ratio (PSNR) and Structure Similarity Index (SSIM).

2 Related Works

Super resolution has a longstanding record of different methods and techniques that have been used in order to upscale the quality of images [5]. The dilemma of simply inflating a low resolution snapshot to construct a compelling elevated rendition is characterized as super resolution. Because of the availability among several strategies of the challenge and the strict preconceived notions that the confluence of imagery must adhere to in order to create convincing high resolution pictures, classic super resolution strategies notably under perform in this realm [6].

Dahl et al. [6] leverage Generative Adversarial Networks (GAN) as its backbone to address the challenge of image enhancement. The implementation of Generative Adversarial Networks has shown excellent outcomes and has great promise for utilization as a viable and sustainable solution which can augment image quality.

Enhanced Super Resolution Generative Adversarial Networks (ESRGAN) proposed by Wang et al. [7] also make use of GANs. Residual-in-Residual Dense Block (RRDB) without batch normalization, which combines multifaceted residual network and dense linkages, is the cornerstone employed for such systems. In order to figure out the chance that the original image is substantially more genuine than a fabricated one, it also incorporates a relativistic discriminator. Relativistic GAN assists the discriminator to anticipate relative "real-ness" rather than an absolute value. It uses network interpolation to reduce interferences in GAN-based approaches while preserving excellent perceptual quality and incorporates the adversarial loss term to the generator loss. Utilizing characteristics that give tighter supervision for recovery of texture and homogeneity of luminosity prior to activation reduces perceptual losses.

A myriad of GAN-based high resolution techniques are built on the Super Resolution Generative Adversarial Network (SRGAN) [4] proposed by Ledig Christian et al. It passes permutations in sequence using numerous residual blocks, skip connections, and up samples before rendering high resolution pictures. It incorporates a perceptual loss function made up of adversarial and pixel loss. Using a discriminator network trained to discern among super-resolved images and authentic photo-realistic imagery, our approach is pushed to the actual picture spectrum by the adversarial loss. Additionally, in lieu of using similarity in the pixel volume, we apply a pixel loss emboldened by perceptual similarity.

Another way with which we can achieve super resolution is by using Deep Convolutional Networks [8–12]. This approach has proven to be quite accurate in the generation of high resolution images from bicubic variants. Super Resolution Convolutional Neural Network is one technique involving use of CNNs (SRCNN) [8]. To get the image to the suitable dimensions, it is first enlarged using bicubic interpolation. Every high-dimensional vector is non linearly translated onto the other after that mapping is implemented to the image in which contiguous regions are excised, and the patches are then amalgamated to form the culminating high resolution visual.

Deep Convolutional Neural Networks (DCNN) proposed by Jin Yamanaka et al. [9] are another technique which makes use of the power of CNNs. This approach utilizes a Convolutional Neural Network with network-in-network and skip connections. With quicker processing and fewer information loss, the reconstruction network's network design achieves improved reconstruction performance.

An improvement of Deep Convolutional Networks [9] is Very Deep Convolutional Networks [10]. The basic rationale behind it is that the model's overall accuracy is vastly enhanced when the depth of the network is expanded by integrating multiple layers. In this research, it is demonstrated that when model depth is increased, model accuracy likewise increases.

The next approaches [13, 14] proposed by Jiang, Yuning and Li, Jinhua and Salvetti Francesco et al. respectively, consist of improvements made to the preexisting models by introducing a new loss term or a new block. Generative Adversarial Networks with Super Resolution similar to ESRGANs, combining texture loss augments

the original SRGAN model. They also include a new texture loss term that is derived from the discriminator's intermediate layers. This loss retrieves feature mappings from the convolutional network of the generator and discriminator network's intermediate layer.

In addition to enhancing the array of data needed to upscale a visual, the Multi-Image Super Resolution issue aggregates images from a variety of timestamps that are incredibly close together in time. Leveraging spatial and temporal correlations, the Residual Attention Multi-image Super Resolution network (RAMS) [14] model is presented in this methodology to merge numerous pictures.

3 Implementation

The aim of the proposed architecture is to achieve image super resolution with improved qualitative results. The architecture of the network is described first, with the improvements introduced over the current state of the art models, and the next subsection elaborates the losses that the model optimizes and their calculation. At last, detail about the training process is provided.

3.1 Dataset

The Div2k dataset [15, 16] consists of RGB images about a variety of topics. The dataset has 1000 high quality images that are divided into 800 train images, 100 validation set images, and 100 test images. The training input low resolution images are taken from the bicubic/X4 folder and the corresponding high resolution images are obtained from HR folder. The model is trained on the entire dataset in order to maximize the learning set.

3.2 Network Architecture

The proposed design for the model is a Generative Adversarial Network which performs image super resolution by generating images based on features learned from input. The model includes a novel generator and a relativistic discriminator. The generator is separated into 4 components that are explained in the following subsections. The architecture is shown in Fig. 1.

Input LR This component comprises of 2 layers namely LR input and a single Conv2d layer to transform the image into desired channels. The main job of this layer is to make sure the input is converted to a shape suitable to the next Feature Attention Network.

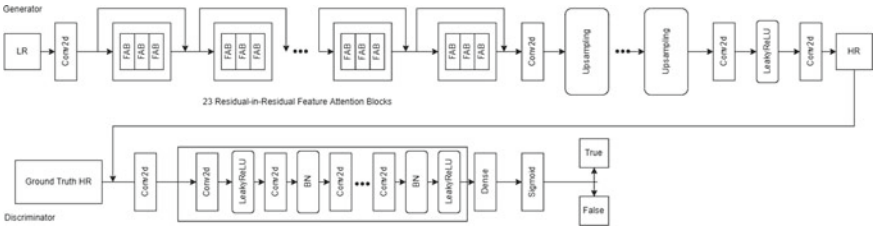

Fig. 1 Model architecture

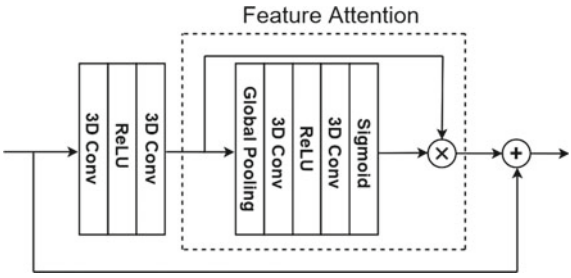

Fig. 2 Feature attention block

Feature Attention Network This section is the most crucial to the network, as it controls which features to look closely at and what features to neglect in order to achieve an accurate high resolution image. This network is built using groups of Feature Attention Blocks. This block is inspired by the Feature Attention Block proposed in [14], where a pair of such blocks are used in parallel to capture features over multiple timesteps and for each local image space.

Feature Attention Blocks have two Conv2d transformations outside them to capture low-level features, which are then passed on to the block itself. The block shown in Fig. 2, contains a global average pooling layer and a sigmoid layer along with the same layers outside it. The intuition here is to capture the high-level features and consequently add the low-level features to them to get a combination of both flowing through the network. A skip connection is also added to allow feature values to pass through unaffected, facilitating a wider variety of functions.

These Feature Attention Blocks are then connected in groups of 3, and a residual connection is added between each group to form the Residual Feature Attention Blocks. The proposed model uses 16 of these blocks in sequence to capture high resolution features from an LR image. The network ends with a single Conv2d layer used to convert the number of channels suitable for the next Upsampling Network.

Upsampling Network The primary job of this component is to convert the information extracted from the Feature Attention Network into a higher resolution image of the desired size. It involves two Transpose Convolutions in the nearest neighbor mode along with 2 convolution layers to get the desired image size as output.

Discriminator The discriminator network uses a classic VGG19 [17] network as basic architecture, which comprises two components: feature extraction and classi-

fier. The feature extraction component has 16 convolutional layers, and each convolution layer has a LeakyReLU activation. The BatchNorm layer is used after each convolutional layer except the first one, to avoid gradient vanishing problems. Instead of a fully connected layer, Global Average Pooling (GAP) is used. GAP calculates the pixel average value across all feature maps, and then all the values are passed to the sigmoid activation function after linear combination. In the end, the network produces D's verdict for the input sample. Generator networks can recover outcomes that are more similar to the original pictures by the outputs of a trained discriminator network.

3.3 Losses

This section expands upon the complex loss function that is used to train FASRGAN. The loss function is given as a weighted sum of Texture Loss L_{tex}, Perceptual Loss L_{per}, Pixel Loss L_{pix}, Adversarial Loss L_{adv}.

$$L = L_{\text{tex}} + \lambda * L_{\text{per}} + \eta * L_{\text{adv}} + L_{\text{pix}}$$

Texture Loss Perceptual loss can enhance the overall quality of the reconstructed picture, but it still introduces extraneous high-frequency features. To create the entire loss function of G, the proposed model includes the texture loss shown in [18]. L_{tex} promotes local texture information matching and retrieves feature maps created by the discriminator network's intermediary layer of a convolutional network. The associated gram matrix is then calculated. The resulting gram matrix values are then utilized to compute texture loss using the L2 loss function:

$$L_{\text{tex}} = ||G(k(I^{\text{gen}}) - G(k(I^{HR}))||_{2'}^2$$

Perceptual Loss L_{per} is called perceptual loss as it is based on VGG network and computes feature layer information before activation layer instead of after it in order to produce pictures that are more visually pleasing. In order to reduce the Euclidean distance between two activation features, it is specified on the activation layer of the pre-trained deep network:

$$L_{\text{per}} = \frac{1}{W_{ij} H_{ij}} \sum_{x=1}^{W_{ij}} \sum_{y=1}^{H_{ij}} \left(k^{ij} \left(I^{HR} \right)_{x,y} - k^{ij} \left(G \left(I^{LR} \right) \right)_{x,y} \right)^2$$

where W_{ij} and H_{ij} represent the shape of the respective feature maps in the VGG model. The improvement overcomes two drawbacks of the original design: The sparsity of activation layer outputs and the inconsistency in brightness of the image.

Pixel Loss To assure the consistency of information between the reconstructed picture and LR image, Mean Square Error (MSE) loss is utilized as the model's pixel loss. Its job is to reduce the squared difference in inaccuracy between pixels in produced and actual HR pictures. Reducing the space between pixels can more rapidly and effectively guarantee the correctness of the information included in the reconstructed image, resulting in greater peak signal-to-noise ratio values. The loss is calculated as:

$$L_{\text{con}} = L_{\text{MSE}}(p) = \frac{1}{N}\sum_{i=1}^{N}\left|\left|I_i^H - G\left(I_i^L, p\right)\right|\right|^2 \quad (1)$$

Adversarial Loss This loss is employed to make the discriminator network produce the probability that an image generated by the generator network is true or false based on the adversarial game mechanism between the two networks. We use the adversarial loss suggested in the WGAN-GP [19] model instead of the one proposed in the GAN model to achieve a probability rather than a discrete class value. Improved L_{adv} punishes D for the input gradient; it can aid in stable GAN architecture training and provide higher-quality samples with quicker convergence times and no need for hyperparameter modification.

$$L_{\text{adv}} = E_{x\sim p_G}[D(x)] - E_{x\sim p_{\text{data}}}[D(x)] + q\,E_{x\sim\text{penalty}}\left[(||d_x D(x)|| - 1)^2\right] \quad (2)$$

4 Results

4.1 Quantitative Analysis

The experiments are performed by enhancing images from the Set5 [20] and Set14 [21] datasets, to observe how introducing the FAB structure improves existing super resolution techniques. For comparison, Bicubic, SRGAN [4], ESRGAN [7], and TSRGAN [13] methods are employed, that were tested on Set5 [20] and Set14 [21]. The average PSNR and SSIM values are delineated in Tables 1 and 2. FASRGAN records a higher PSNR and SSIM value when compared to all other methods and indicates a better performance as compared to existing methods.

The reconstruction time for the different techniques is also compared as shown in Table 3. As it can be seen, the significant improvement in PSNR and SSIM scores comes without much compromise on speed, verifying its effectiveness.

4.2 Qualitative Analysis

To ensure the correctness of the reconstructed images, a qualitative evaluation is performed by observing the reconstruction results. The reconstructed images from

Table 1 PSNR and SSIM values on Set5 [20]

Algorithm	PSNR	SSIM
Bicubic	30.07	0.862
SRGAN	30.36	0.873
ESRGAN	32.05	0.895
TSRGAN	32.38	0.967
FASRGAN	31.07	0.975

Table 2 Reconstruction times on Set14 [21]

Algorithm	PSNR	SSSIM
Bicubic	27.18	0.786
SRGAN	27.02	0.772
ESRGAN	28.49	0.819
TSRGAN	28.73	0.810
FASRGAN	31.21	0.9578

Table 3 Reconstruction times on Set5 [20] and Set14 [21]

Algorithm	Set5[20]	Set14[21]
Bicubic	1.725	1.816
SRGAN	3.763	4.098
ESRGAN	3.247	3.862
TSRGAN	3.750	3.899
FASRGAN	2.770	2.885

the Bicubic, SRGAN, ESRGAN, and TSRGAN techniques are shown in Fig. 3 along with the original high resolution image. The effectiveness of the proposed technique is made evident when we compare the resulting images.

As it can be observed, FASRGAN better resolves the finer details of images as compared to their counterparts, hence proving its qualitative superiority. Edges are well-contained in the outputs without affecting the areas around it. The result is also a smooth image that preserves the major as well as minor features.

5 Conclusion

The proposed model is a successful novel approach for Single Image Super Resolution. The new model is called FASRGAN and is built by improving the network architecture of existing GAN-based SR techniques. It is evident from the quantitative and qualitative analysis that FASRGAN outperforms the techniques that inspired

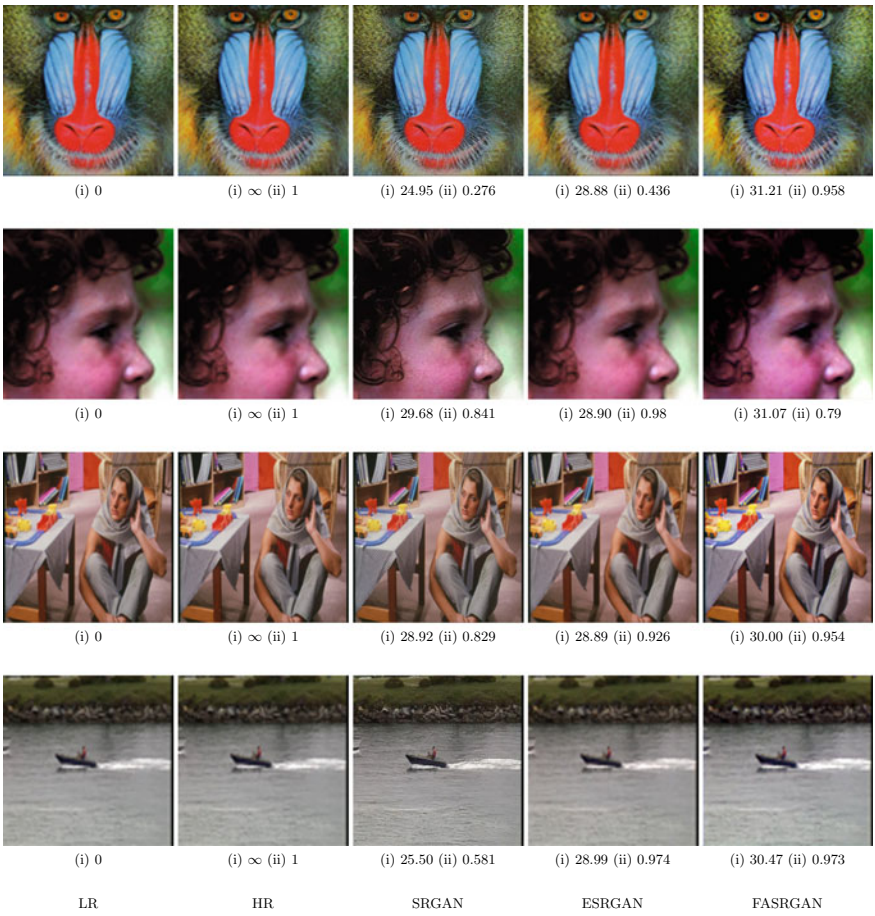

Fig. 3 The results of X4 for LR, HR, SRGAN, ESRGAN, and the proposed FASRGAN on Set5[20] and Set14[21] images (Note: (i),(ii) in Fig. 3 stand for the PSNR and SSIM values of the image, respectively)

it. The model currently uses three losses—Pixel (MSE) Loss, Adversarial Loss, and Perceptual Loss. The model however is not perfect and has room for further improvement.

The future scope of this technique includes the ability to improve on color correction ability. The model faces a setback when predicting the pixels of a highly dynamic portion of an image, which includes a region of an image which contains pixels occupied by edges of multiple objects. A model that can improve upon this while maintaining or upgrading the super resolution quality is most desirable.

References

1. Schultz R, Stevenson R (1994) A Bayesian approach to image expansion for improved definition. IEEE Trans Image Process 3(3):233–242
2. Gribbon KT, Bailey DG (2004) A novel approach to real-time bilinear interpolation. In: Proceedings. DELTA 2004. Second IEEE international workshop on electronic design, test and applications, pp 126–131
3. Zhang L, Wu X (2006) An edge-guided image interpolation algorithm via directional filtering and data fusion. IEEE Trans Image Process 15(8):2226–2238
4. Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, Shi W (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 105–114
5. Mittal A, Soundararajan R, Bovik AC (2013) Making a “completely blind" image quality analyzer. IEEE Signal Process Lett 20(3):209–212
6. Dahl R, Norouzi M, Shlens J (2017) Pixel recursive super resolution. CoRR, arXiv:abs/1702.00783
7. Wang X, Yu K, Wu S, Gu J, Liu Y, Dong C, Loy CC, Qiao Y, Tang X (2018) ESRGAN: enhanced super-resolution generative adversarial networks. CoRR, arXiv:abs/1809.00219
8. Dong C, Loy CC, He K, Tang X (2016) Image super-resolution using deep convolutional networks. IEEE Trans Pattern Anal Mach Intell 38(2):295–307
9. Yamanaka J, Kuwashima S, Kurita T (2017) Fast and accurate image super resolution by deep CNN with skip connection and network in network. CoRR, arXiv:abs/1707.05425
10. Kim J, Lee JK, Lee KM (2016) Accurate image super-resolution using very deep convolutional networks. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp 1646–1654
11. Cai J, Zeng H, Yong H, Cao Z, Zhang L (2019) Toward real-world single image super-resolution: a new benchmark and A new model. CoRR, arXiv:abs/1904.00523
12. Gohshi S (2015) Real-time super resolution algorithm for security cameras. In: 12th international joint conference on e-business and telecommunications (ICETE), vol 05, pp 92–97
13. Jiang Y, Li J (2020) Generative adversarial network for image super-resolution combining texture loss. Appl Sci 10(5)
14. Salvetti F, Mazzia V, Khaliq A, Chiaberge M (2020) Multi-image super resolution of remotely sensed images using residual attention deep neural networks. Remote Sens 12(14)
15. Agustsson E, Timofte R (2017) Ntire 2017 challenge on single image super-resolution: dataset and study. In: The IEEE conference on computer vision and pattern recognition (CVPR) workshops, July 2017
16. Timofte R, Agustsson E, Van Gool L, Yang M-H, Zhang L, Lim B et al (2017) Ntire 2017 challenge on single image super-resolution: methods and results. In: The IEEE conference on computer vision and pattern recognition (CVPR) workshops, July 2017
17. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
18. Sajjadi MSM, Schölkopf B, Hirsch M (2016) Enhancenet: single image super-resolution through automated texture synthesis. CoRR, arXiv:abs/1612.07919
19. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC (2017) Improved training of Wasserstein GANS. CoRR, arXiv:abs/1704.00028
20. Bevilacqua M, Roumy A, Guillemot C, Line Alberi Morel M (2012) Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: Proceedings of the British machine vision conference. BMVA Press, pp 135.1–135.10
21. Zeyde R, Elad M, Protter M (2010) On single image scale-up using sparse-representations. In: Proceedings of the 7th international conference on curves and surfaces. Springer-Verlag, Berlin, Heidelberg, pp 711–730