Single Image Super-Resolution Using a Generative Adversarial Network

V Ramya *ee21btech11058*

V Pooja *ee21btech11059*

G Vandana *ee21btech11019*

K Moulika *ee21btech11028*

N Sanskruti *ee21btech11036*

***Abstract*—Despite advancements in speed and accuracy of single image super-resolution with deeper convolutional neural networks, recovering finer texture details at large upscaling factors remains a challenge. Optimization-based methods primarily minimize mean squared reconstruction error, resulting in high peak signal-to-noise ratios but lacking high-frequency details. We introduce SRGAN, a Generative Adversarial Network (GAN), capable of inferring photo-realistic natural images at 4× upscaling factors. Our approach incorporates a perceptual loss function comprising adversarial and content losses. The discriminator network guides solutions towards the natural image manifold, while the content loss emphasizes perceptual similarity. In this paper we will see that SRGAN achieves significant gains in perceptual quality.**

I. Introduction

Image super-resolution involves enlarging small images while minimizing the loss in quality, or restoring high-resolution images from detailed information extracted from low-resolution counterparts. The complexity arises from multiple potential solutions for a given low-resolution image. This technique is used in various fields such as satellite and aerial image analysis, medical image processing, and enhancing compressed image or video content.

In 2014, Dong et al. introduced convolutional neural networks (CNNs) to image super-resolution with Super-Resolution using Convolutional Neural Network (SRCNN). This uses a 3-layer CNN structure to learn correlations between low and high-resolution images, resulting in improved reconstructions. However, the shallow hierarchy of the 3-layer network limits its ability to capture deep image features. Recently, Residual networks (ResNets) have been employed to address this limitation, leveraging increased network depth to enhance image quality in super-resolution tasks.

Here we have a Super-Resolution Generative Adversarial Network (SRGAN), employing a deep residual network (ResNet) with skip-connections, diverging from Mean Squared Error (MSE) as the optimization target. This method for perceptual loss utilizes high-level feature maps from the VGG network and a discriminator to encourage solutions resembling the high-resolution reference images and are challenging to distinguish from them.

II. Problem Statement

The objective is to recover or restore high-resolution images from their low-resolution counterparts. Image enhancement encompasses many techniques such as noise reduction, upscaling, and color adjustments. This paper focuses on enhancing low-resolution images by applying deep networks with adversarial networks (Generative Adversarial Networks) to produce high-resolution images.

The aim is to recontruct super-resolution images or high-resolution images by up-scaling low-resolution images such that texture detail in the reconstructed SR images is preserved.

III. literature Review

There exist several methods for improving image quality, with interpolation being one of the most frequently employed techniques. While interpolation is straightforward to apply, it often results in image distortion or a reduction in visual quality. Conventional interpolation methods, such as bi-cubic interpolation, typically yield blurry images. However, more advanced techniques leverage internal similarities within an image or utilize datasets containing pairs of low-resolution and high-resolution images to learn an effective mapping between them. Within the realm of Example-Based Super-Resolution (SR) algorithms, the Sparse-Coding-Based method stands out as one of the most widely utilized approaches.

Deep learning offers a superior approach for obtaining optimized images, particularly in the realm of Image Super Resolution (ISR). Over recent years, numerous methods have been introduced to address this task. One prominent technique we will delve into is SRGAN (Super-Resolution Generative Adversarial Network). Additionally, there exist various other methodologies within the domain of deep learning for image enhancement and super-resolution:

- SRCNN : SRCNN (Super-Resolution Convolutional Neural Network) was a groundbreaking deep learning approach that surpassed traditional methods in image enhancement. It comprises a Convolutional Neural Network with just three convolutional layers: patch extraction and representation, non-linear mapping and reconstruction.
- SRGAN : A novel method introduced by Ledig [1] in image super-resolution utilizes Generative Adversarial Networks [2] (GANs). This technique involves a generator that produces high-resolution images, refined by a discriminator through optimization. The conventional use of Mean Squared Error (MSE) as a loss function often results in images with smoothed edges and lacking fine details. To address this, a perceptual loss function, denoted as lSR, is proposed, comprising both content loss and adversarial loss components. This approach is designed to better align with human visual perception, thereby improving the realism and quality of the generated images.

IV. SRGAN — Super-Resolution Generative Adversarial Network

GANs represent a class of AI algorithms utilized in unsupervised machine learning. The GAN framework encourages reconstructions to gravitate towards regions within the search space that exhibit a high probability of containing photo-realistic images, thereby bringing them closer to the natural image manifold.

GANs consist of a Generator and Discriminator. Discriminator decides whether input is coming from true training data set of fake generated data. In training, I^{LR} is obtained by applying a Gaussian filter to I^{HR} followed by a downsampling operation with downsampling factor r. I^{LR} is the low-resolution version of high resolution version I^{HR}.Generator tries to optimize data so that it can match true training data.

Our ultimate goal is to train a generating function G that estimates for a given LR input image its corresponding HR counterpart. To achieve this, we train a generator network as a feed-forward CNN G_{θ_G} parametrized by θ_G. Here θ_G =$W_{1:L}$;$b_{1:L}$ denotes the weights and biases of a L-layer deep network and is obtained by optimizing a SR-specific loss function l^{SR}. For training images I_n^{HR}, n=1,...N with corresponding I_n^{HR}, n=1,...N,we solve:

$$\hat{\theta}_{\mathrm{G}} = \arg\min_{\theta_G} \frac{1}{N} \sum_{n=1}^{N} l^{SR}(G_{\theta_G}(I_n^{LR}), I_n^{HR})(1)$$

In this work we will specifically design a perceptual loss I^{SR} as a weighted combination of several loss components that model distinct desirable characteristics of the recovered SR image.

A. Adversarial network architechture

We further define a discriminator network D_{θ_D} which we optimize in an alternating manner along with G_{θ_G}

$$\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{I^{\mathrm{HR}} \sim p_{\mathrm{train}}(I^{\mathrm{HR}})}[\log D_{\theta_D}(I^{\mathrm{HR}})]+ \quad (2)$$

$$\mathbb{E}_{I^{\mathrm{LR}} \sim p_G(I^{\mathrm{LR}})}[\log(1 - D_{\theta_D}(G_{\theta_G}(I^{\mathrm{LR}})))]$$

This methodology trains a generative model G to deceive a discriminator D, distinguishing super-resolved images from real ones. Our generator employs two convolutional layers with 3x3 kernels and 64 feature maps, followed by batch-normalization layers and ParametricReLU. Additionally, we increase input image resolution using two trained sub-pixel convolution layers. To discern real high-resolution (HR) images from generated SR samples, we utilize a discriminator network with LeakyReLU activation (α = 0.2) and no max-pooling layers.

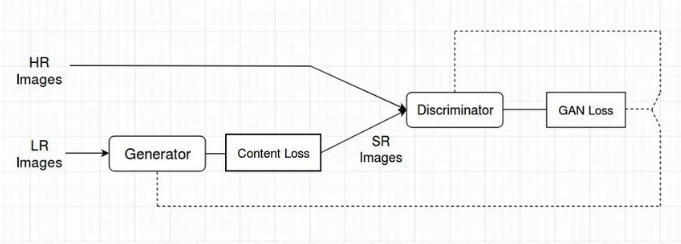


Fig. 1. **SRGAN Architecture**

B. Perceptual loss function

The definition of our perceptual loss function, denoted as l^{SR}, plays a pivotal role in determining the effectiveness of our generator network. While l^{SR} is often constructed using the Mean Squared Error (MSE) [8] as a model, we refine upon the works of Bruna et al [4]. and devise a loss function that evaluates a solution based on perceptually significant attributes. Our formulation of the perceptual loss involves a weighted combination of a content loss (l_X^{SR}) and an adversarial loss component as follows:

$$l^{SR} = l_X^{SR} + 10^{-3} l_{Gen}^{SR} \tag{4}$$

In the following, we describe possible choices for the content loss l_X^{SR} and the adversarial l_{Gen}^{SR}.

C. Content loss

$$l_{MSE}^{SR} = \frac{1}{r^2WH}\sum_{x=1}^{rW}\sum_{y=1}^{rH}(I_{x,y}^{HR} - G_{\theta_G}(I^{LR})_{x,y})^2 \tag{5}$$

However, while achieving notably high PSNR (Peak Signal-to-Noise Ratio), solutions derived from MSE (Mean Squared Error) optimization often lack high-frequency content, resulting in visually unsatisfactory outcomes with overly smooth textures. Hence, we opt for a loss function that prioritizes perceptual similarity. Our approach involves defining the VGG loss, which leverages the ReLU activation layers of a pre-trained VGG19 network comprising 19 layers. Using indices i and j, we denote the feature map obtained by the j-th convolution (post-activation) before the i-th max-pooling layer within the VGG19 network, which remains fixed. Consequently, the VGG loss is established as the Euclidean distance between the feature representations of a reconstructed image $G_{\theta_G}(I^{LR})$ and the reference image I^{HR}.

$$l_{VGG/i,j}^{SR} = \frac{1}{W_{i,j}H_{i,j}}\sum_{x=1}^{W_{i,j}}\sum_{y=1}^{H_{i,j}}(\Phi_{i,j}(I^{HR})_{x,y} - \Phi_{i,j}(G_{\theta_G}(I^{LR}))_{x,y})^2 \tag{6}$$

Here $W_{i,j}$ and $H_{i,j}$ describe the dimensions of the respective feature maps within the VGG network

D. Adversarial loss

We also add the generative component of our GAN to the perceptual loss.The generative loss I_{Gen}^{SR} is defined based on the probabilities of the discriminator $D_{\theta_D}(G_{\theta_G}(I^{LR}))$ over all training samples as:

$$l_{Gen}^{SR} = \sum_{n=1}^{N} -\log D_{\theta_D}(G_{\theta_G}(I^{\mathrm{LR}})) \tag{8}$$

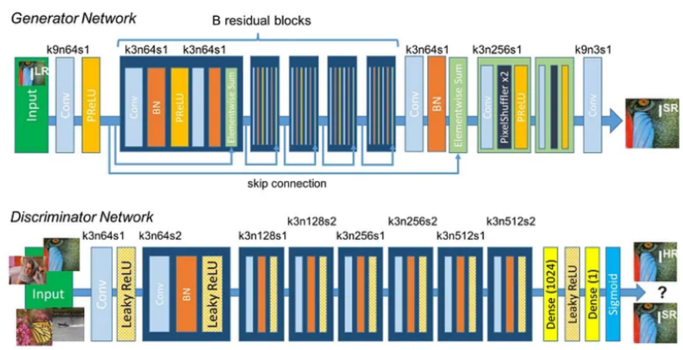


Fig. 2. **Generative and Adversarial Network**

V. Experiments

A. Data and Similarity Measures

We conduct experiments using three commonly employed benchmark datasets: Set5 [5], Set14 [6], and BSD100 , which is the testing set of BSD300 [7]. In all experiments, we maintain a scale factor of 4× between low- and high-resolution images, resulting in a significant 16× reduction in image pixels.Super-resolved images are generated by methods, such as bicubic interpolation, SRGAN in this project To compare and evaluate our results we use the following evaluation metrics:

(a)PSNR (Peak Signal-to-Noise Ratio): It is the ratio between the maximum power produced by the signal and the noise that distorts the image's actual

representation. It is typically expressed in decibels (dB).

(b)SSIM (Structural Similarity Index Measure): It gives the measure of similarity between two images, ranging from -1 to 1. It provides a comparative measure of similarity between two pixels are, considering their structural properties.

(c)MOS (Mean Opinion Score): It is a rating provided by evaluators, ranging from 1 (lowest) to 5 (highest).

B. Bicubic Interpolation

Bicubic interpolation is a technique commonly employed in image processing to enhance the resolution of images while maintaining smoothness and reducing visual artifacts. In the provided code, the process begins by loading a high-resolution reference image.The low-resolution (LR) images were obtained by downsampling the high-resolution (HR) images (BGR, C = 3) using a bicubic kernel with a downsampling factor of r = 4. This lower resolution image serves as the input for the subsequent upscaling step. Bicubic interpolation is again applied to enlarge the low-resolution image, effectively increasing its dimensions to match those of the original high-resolution image.This method involves examining a 4x4 grid surrounding each new pixel to estimate its value.Subsequently, a weighted average is computed and the pixel values of the neighboring pixels, where closer pixels are given higher weights based on their distance from the target position.Upon calculating the new pixel values for all positions in the resized image, a higher-resolution version of the lower-resolution counterpart image is achieved.To assess the quality of the obtained upscaled image, various image quality metrics such as Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), are calculated by comparing it with the original high-resolution image.

C. Mean Opinion Score (MOS) Testing

We perform a MOS (Mean Opinion Score) test to quantify the ability of different approaches to reconstruct perceptually convincing images. Specifically, all the members of our group assigned an integral score from 1 (bad quality) to 5 (excellent quality) to the super-resolved images. We rated the bicubic

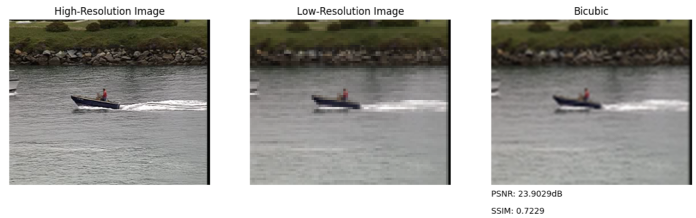


Fig. 3. **Original HR image, LR image, Bicubic interpolation image and corresponding PSNR and SSIM index are shown below[4 upscaling].**

interpolated versions of each image in the data sets Set5, Set14 and BSD100. We observed high reliability and noted no significant variations in ratings for identical images. The experimental results of the conducted MOS tests, PSNR and SSIM Index are summarized in the table given below:

TABLE I
BICUBIC

Bicubic	Set5	Set14	BSD100
PSNR	26.51	24.18	25.41
SSIM	0.932	0.842	0.86
MOS	1.72	1.67	1.31

D. SRGAN Implementation

This architecture follows a GAN-based approach, which means it incorporates both a generator and a discriminator. The generator aims to produce photorealistic images, while the discriminator is tasked with distinguishing the generated images and the original ones. But first we need to do some data preprocessing.

1) Data Preprocessing:

To handle image data, we create functions which 1) converts a list of images into a high-resolution numpy array; 2) generates low-resolution images from high-resolution images by downscaling them using bicubic interpolation.

2) Generator Network:

The generator network is a feedforward CNN designed to convert Low Resolution images into High Resolution images using a specific loss function tailored for Super-Resolution (SR). It comprises 16 Residual blocks and 2 upsampling blocks, each serving distinct purposes in the image transformation process.

(a)Convolutional Block: This block executes convolutions to extract features from the input noise, with 64 filters of size 3x3 to extract features from the input image.
(b)Batch Normalization: Following feature extraction, batch normalization is applied to normalize the dataset, reducing computation and preventing internal covariate shifts.
(c)Leaky ReLU: The generator employs Leaky ReLU as the activation function, to deal with negative values and avert the dead neuron problem. Leaky ReLU multiplies negative results by a parameter (alpha = 0.02), enhancing the network's capacity to learn complex features.
(d)Convolutional Layer with Batch Normalization: After Leaky ReLU layer, another convolutional layer is applied, followed by batch normalization, to further refine the features extracted from the input.
(e)Elementwise Sum Method: Finally, an elementwise sum operation is performed, integrating the features extracted from previous layers and preserving essential image details.

3) *Discriminator Network:*

The discriminator network architecture employs convolutional layers paired with the Leaky ReLU activation function, featuring an alpha value set to 0.2. The initial block of the discriminator architecture integrates a convolutional layer followed by a Leaky ReLU activation function, while the subsequent five blocks consist of a convolutional layer, followed by batch normalization, and culminating with an additional Leaky ReLU activation function layer. The final layers of the architecture include fully connected nodes with parameters, a Leaky ReLU layer, and a final fully connected dense node utilizing the sigmoid activation function for classification purposes.
(a)Convolutional Block: This conducts convolutions with a filter size of 3x3 of each block to extract features.
(b)Batch Normalization: This layer aims to normalize the features extracted by the discriminator, to reduce computational complexity.
(c)LeakyReLU Activation Function: LeakyReLU is used to introduce a small slope for negative values, to help prevent the zero slope error, ensuring negative values to be taken into consideration.
(d)Dense Block: This is used to convert and connect tensor feature maps back to a 1D array.
(e)Sigmoid Function: As the final classifier, the sigmoid function is used to classify the output as either High Resolution (HR) or Super-Resolution (SR). It converts the tensors into a logical number, that shows the probability of the image being HR or SR.

VI. Results

We compare the performance of SRGAN to bicubic interpolation and observe that SRGAN outperforms bicubic interpolation by a large margin and sets a new state of the art for photo-realistic image SR. After obtaining MOS ratings on the Dataset BSD100, we also see that all the differences in MOS are highly significant for SRGAN compared to bicubic interpolation.

TABLE II
SRGAN

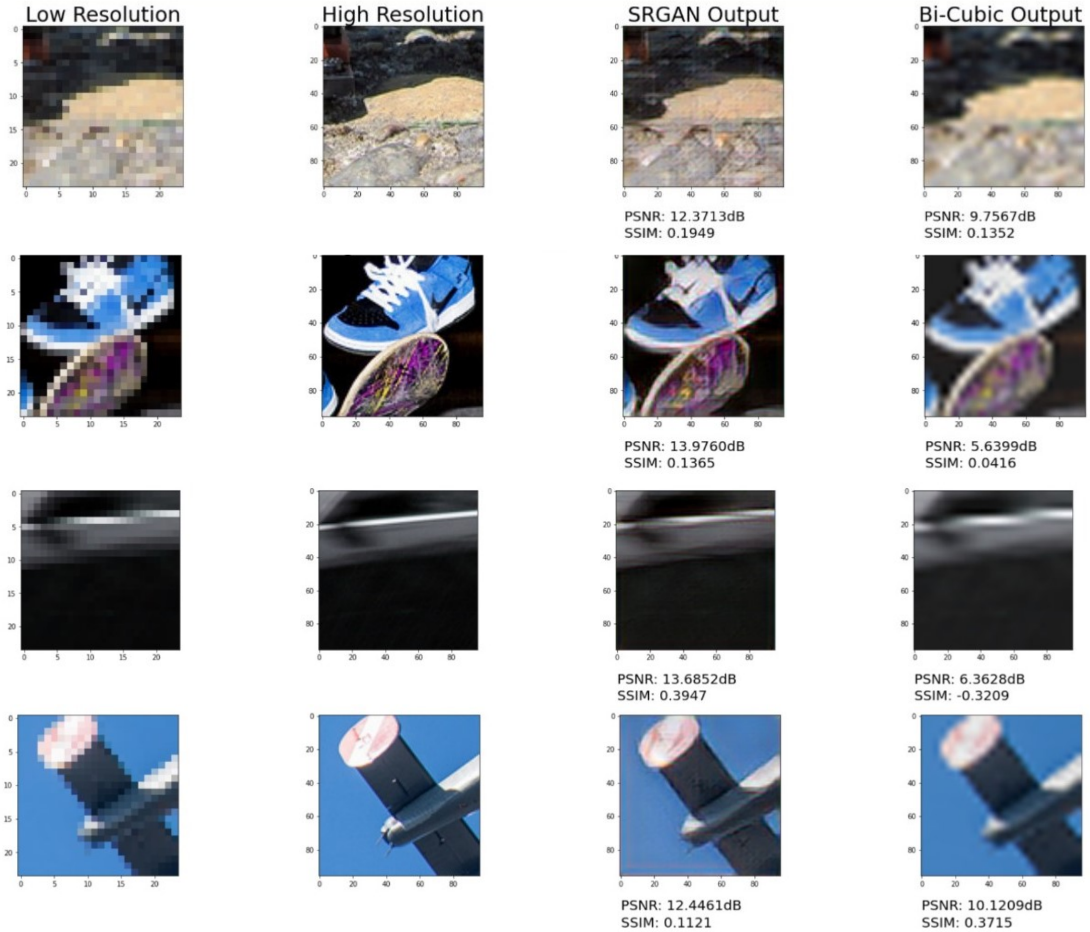
SRGAN	Set5	Set14	BSD100
PSNR	12.48	13.66	12.55
SSIM	0.34	0.28	0.31
MOS	3.12	3.34	2.98

VII. Conclusion and Future Work

We employed an SRGAN model, which augments the content loss function with an adversarial loss by training a GAN. Using extensive MOS testing, we have confirmed that SRGAN reconstructions for large upscaling factors (4×) are more photo-realistic by a considerable margin.

We also found that standard quantitative measures like PSNR and SSIM don't accurately reflect image quality as perceived by humans. Our focus was on the perceptual quality of super-resolved images rather than computational efficiency. Unlike Shi et al. [9], our model isn't optimized for real-time video SR. However, initial experiments suggest that shallower networks could offer efficient alternatives with slightly reduced quality. In contrast to Dong et al. [8], we observed deeper network architectures to be advantageous.

We note that the ideal loss function varies depending on the application. For instance, methods that generate fine details might not be suitable for medical or



surveillance purposes. The accurate reconstruction of text or structured scenes remains challenging and can be a part of future research. Additionally, developing content loss functions that describe image spatial content but are more invariant to pixel space changes will further enhance photo-realistic image super-resolution results.

References

[1] C. Ledig et al., "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 105-114, doi: 10.1109/CVPR.2017.19.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., "Generative adversarial nets", Advances in Neural Information Processing Systems (NIPS), pp. 2672-2680, 2014.

[3] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(2):295–307, 2016.

[4] J. Bruna, P. Sprechmann, and Y. LeCun. Super-resolution with deep convolutional sufficient statistics. In International Conference on Learning Representations (ICLR), 2016. 2, 3, 5

[5] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel.Low-complexity single-image super-resolution based on nonnegative neighbor embedding. BMVC, 2012

[6] R. Zeyde, M. Elad, and M. Protter. On single image scale-up using sparse-representations. In Curves and Surfaces, pages 711–730. Springer, 2012

[7] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In IEEE International Conference on Computer Vision (ICCV), volume 2, pages 416–423, 2001

[8] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(2):295 307, 2016. 3, 5, 7

[9] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1874–1883, 2016