Exploring Image Super Resolution Using SRGAN

Amogh Dattadas Kubal
New York University
New York, USA
ak10499@nyu.edu

Pranav Thorat
New York University
New York, USA
pat9991@nyu.edu

***Abstract*—Building upon the remarkable strides in single image super-resolution, our project implements and examines the capabilities of SRGAN, a state-of-the-art generative adversarial network. While progress has been made in the efficiency and effectiveness of convolutional neural networks in this domain, achieving finely textured details at high upscaling factors remains a challenge. SRGAN addresses this by utilizing a perceptual loss function, comprising adversarial and content loss components. The adversarial loss enables the network to generate images that lie on the natural image manifold, by leveraging a discriminator network that distinguishes between super-resolved and authentic photo-realistic images. The content loss is influenced by perceptual similarity rather than pixel accuracy, aligning more closely with human visual perception.**

In our project, we replicate the SRGAN framework, aiming to understand its intricacies and to explore its potential for 4X upscaling of images. We delve into the neural network's ability to recover photo-realistic textures from significantly downsampled images. Our endeavor does not extend to new methodologies or improvements but serves as an academic exercise to validate the reproducibility of the SRGAN's performance on public benchmarks. This replication also provides insights into the practical aspects of deploying such deep learning models and their behavior under different conditions and parameters. Our findings reiterate the efficacy of the SRGAN model in super-resolution tasks, emphasizing the importance of perceptual congruity in the realm of image upscaling.

I. Introduction

Super-resolution (SR) is the demanding process of generating a high-resolution (HR) image from a low-resolution (LR) one. This task has garnered significant interest among researchers in the field of computer vision and is applied in various contexts [2, 3, 4]

Keeping this in mind, we aim to implement and examine the capabilities of SRGAN [1], a state of the art generative adversarial network that attempts to incorporate the finer texture details into constructing super resolved images.

In our study, we have learned from the referenced research paper and other cited works that the super-resolution (SR) problem becomes more complex with higher upscaling factors, often resulting in a lack of texture detail in the enhanced images. It's noted that supervised SR algorithms typically aim to minimize the mean squared error (MSE) between the enhanced high-resolution image and the original. This approach is advantageous as it aligns with maximizing the peak signal-to-noise ratio (PSNR), a prevalent metric for assessing SR algorithms [5]. However, we observed that both MSE and PSNR have limitations in accurately reflecting perceptual differences, such as texture details, due to their reliance on pixel-wise comparisons [6, 7, 8]. Christian Ledig, Lucas Theis, and colleagues introduced a super-resolution generative adversarial network (SRGAN) that utilizes a deep residual network (ResNet) with skip connections, moving away from the exclusive use of mean squared error (MSE) for optimization. Uniquely, their approach incorporates a new perceptual loss based on the high-level feature maps from the VGG network [9, 10, 11], along with a discriminator designed to generate solutions that are perceptually indistinguishable from high-resolution reference images.

II. Dataset

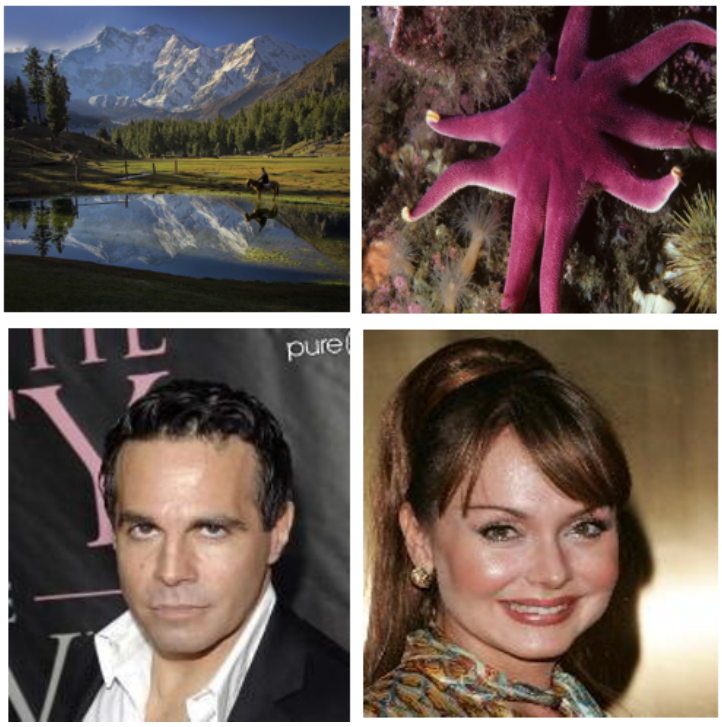

Fig. 1. Sample from CelebA and DIV2K

We have trained our network on 2 different datasets, the DIV2K dataset [12] and CelebFaces Attributes Dataset (CelebA). CelebFaces Attributes Dataset (CelebA) is a large-scale face attributes dataset with more than 200K celebrity images, each with 40 attribute annotations. The images in

this dataset cover large pose variations and background clutter. CelebA has large diversities, large quantities, and rich annotations, including 10,177 number of identities, 202,599 number of face images, and 5 landmark locations, 40 binary attributes annotations per image.However, we have used a subset of 1000 images for training purpose. The DIV2K dataset consists of 1,000 high-resolution images. These images are carefully curated to cover a wide range of subjects and scenarios, ensuring a diverse and representative sample for super-resolution tasks. Each image in the dataset is of 2K resolution, providing detailed and high-quality ground truth for super-resolution models. This high resolution is crucial for evaluating the effectiveness of super-resolution techniques in recovering fine details and textures. DIV2K is systematically divided into distinct subsets: 800 images for training, 100 for validation, and 100 for testing. This structure facilitates a standardized approach to training and benchmarking super-resolution models. The dataset includes low-resolution counterparts for each high-resolution image, obtained through downscaling at different factors, typically x2, x3, and x4. This variation allows for the training and testing of models across multiple scales, a critical aspect in assessing the robustness and versatility of super-resolution algorithms.

III. Methodology

Drawing on the foundational research presented in [1] the objective of Single Image Super-Resolution (SISR) is to infer a super-resolved image I^{SR} from a given low-resolution (LR) input I^{LR}. The LR image is a scaled-down version of its high-resolution (HR) counterpart I^{HR} which is only accessible during the training phase. Typically, I^{LR} is generated by subjecting I^{HR} to a Gaussian blur, followed by downscaling it by a factor r. These images are represented as real-valued tensors of size W X H X C for the LR images and rW X rH X C for the HR images, where C denotes the number of color channels. We adopted the following equation for loss in our implementation[1]

$$\hat{\theta}_G = \arg\min_{\theta_G} \frac{1}{N} \sum_{n=1}^{N} L^{SR}\left(G_{\theta_G}(I_n^{LR}), I_n^{HR}\right)$$

We use SRGAN which is the GAN-based network optimized for a new perceptual loss. Here we replace the MSE-based content loss with a loss calculated on feature maps of the VGG network [13], which are more invariant to changes in pixel space

A. Architecture

At the core of our very deep generator network G, which is illustrated in Figure below are B residual blocks (16) with identical layout. Specifically, we use two convolutional layers with small 3×3 kernels and 64 feature maps followed by batch-normalization layers and Parametric-ReLU as the activation function. The residual block is preceded and succeeded b 2 conv layers of kernel 9X9 and 3x3 respectively. The first conv layer followed by activation P-ReLU and the second one simply have a 2D Batch Normalization appended. This

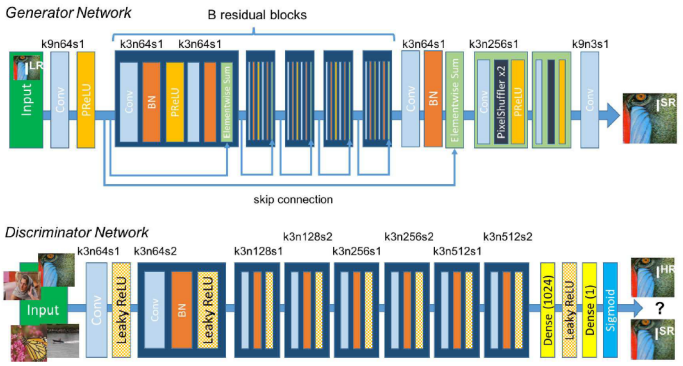

Fig. 2. Architecture of SRGAN

is followed by 2 up sampling layers, using which we increase the resolution of the input image with these two trained sub-pixel convolution layers. This followed by a final Conv layer of 64 input channels, 3 output channels and a kernel size of 9x9. To discriminate real HR images from generated SR samples we train a discriminator network. The architecture is shown in Figure 2. We use LeakyReLU activation $(= 0.2)$ and avoid max-pooling throughout the network. It contains eight convolutional layers with an increasing number of 3 × 3 filter kernels, increasing by a factor of 2 from 64 to 512 kernels as in the VGG [13] network. Strided convolutions are used to reduce the image resolution each time the number of features is doubled. The resulting 512 feature maps are followed by two dense layers and a final sigmoid activation function to obtain a probability for sample classification.

B. Data Preprocessing

The images from the dataset are initially resized to a resolution of 256*256 and then the tensor image is normalized with mean and standard deviation. These are high quality images. The corresponding low quality images are generated by down-scaling the above images by a factor of 4, followed by a similar image tensor normalization.

C. Loss

We are using the pre-trained model – VGG19, for feature extraction. While training the generator, the content loss (criterion L1 loss) is calculated using the features extracted by the above pre-trained model on real HQ images and the corresponding HQ images produced by the generator using LQ images passed to it. While training the discriminator, the adversarial loss (using MSE) is calculated in 2 parts. First one calculates loss based on the "real" or the dataset HQ images, whereas the second loss is calculated based on the "fake" or the images produced by the generator

- Perceptual loss function : The definition of the perceptual loss function l^{SR} is critical for the performance of the generator network. While l^{SR} is commonly modeled based on the MSE [14, 15] and design a loss function that assesses a solution with respect to perceptually

relevant characteristics. We formulate the perceptual loss as the weighted sum of a content loss l_X^{SR} and an adversarial loss component as:

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

We use the below loss function [1],

$$l_{VGG_{i,j}}^{SR} = \frac{1}{W_{i,j} H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I^{HR})_{x,y} -$$

$$\phi_{i,j}(G_{\theta_G}(I^{LR}))_{x,y}^2$$

The feature maps can be seen in Figure 3, Figure 4, Figure 5 and Figure 6.

- Pixel-wise loss functions such as MSE struggle to handle the uncertainty inherent in recovering lost high-frequency details such as texture: minimizing MSE encourages finding pixel-wise averages of plausible solutions which are typically overly-smooth and thus have poor perceptual quality. Reconstructions of varying perceptual quality are exemplified with corresponding PSNR. We illustrate the problem of minimizing MSE in Figure 3 where multiple potential solutions with high texture details are averaged to create a smooth reconstruction.

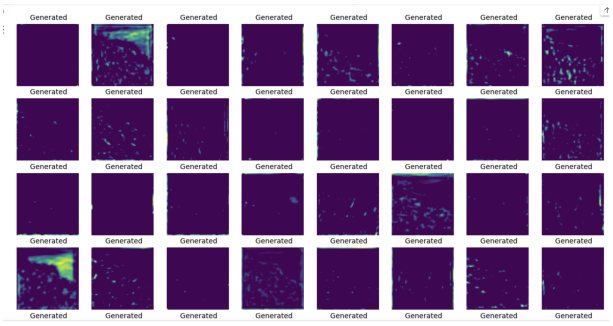

Fig. 3. Feature Map from Generated Images

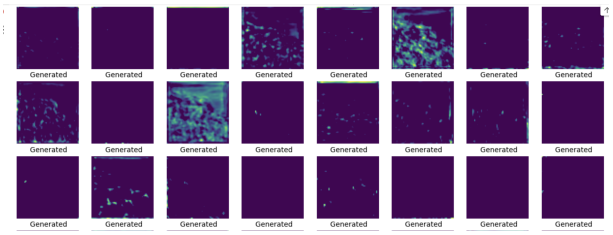

Fig. 4. Feature Map from Generated Images

IV. Experiments

In our experiments, the Generative Adversarial Network (GAN) was subjected to training using two distinct datasets:

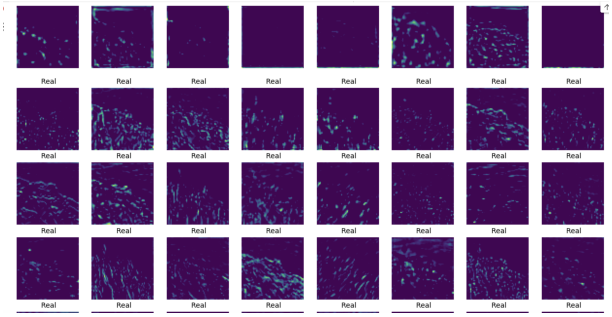

Fig. 5. Feature Map from Real Images

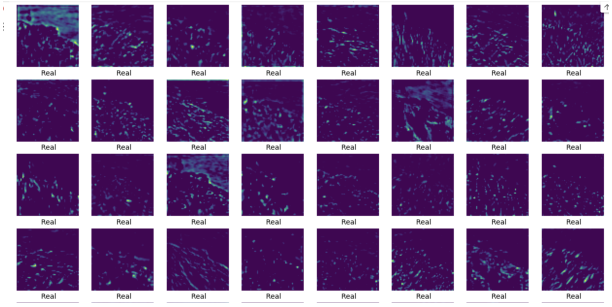

Fig. 6. Feature Map from Real Images

CelebA and DIV2K. This process successfully yielded super-resolved images from our generator with both datasets. Through iterative testing, we ascertained that a batch size of 4, coupled with a learning rate of 0.0002, facilitated the most efficient training progression in terms of loss metrics. Alternative configurations, including larger batch sizes such as 24 and varying learning rates of 0.0001 and 0.0005, were also evaluated. To monitor and save progress, we instituted a checkpoint system that saved the states of the generator and discriminator every 40 epochs.

Originally, our network underwent training over 200 epochs; however, this extensive training led to overfitting. Adjustments were made by curtailing the training to 120 epochs, which empirically emerged as the most effective duration to prevent overfitting while ensuring adequate learning by the network.

V. Results

The presented images showcase the results of our implementation of the Super-Resolution Generative Adversarial Network (SRGAN) [1] . Each pair of images demonstrates the transformative ability of SRGAN, with the left side depicting the original low-resolution input and the right side exhibiting the super-resolved output. From the visual evidence, it is apparent that our SRGAN model has significantly enhanced the resolution of the input images, restoring a considerable amount of detail and sharpness. For instance, in the case of the basketball players in Figure 8, the SRGAN output shows clearer lines and contours, as well as improved texture on the court surface and the audience in the background. Similarly, the butterfly wings exhibit more intricate patterns, and the

Fig. 7. Results

Fig. 8. Results

facial features of the individuals are more distinct in figures 8 and 9 in the super-resolved images.

The SRGAN model's performance can be attributed to its deep learning architecture, which captures and reconstructs high-frequency details that are typically lost in lower-resolution images. The generator's use of residual blocks promotes the learning of intricate details, while the discriminator ensures that the generated images are indistinguishable from high-resolution references. The perceptual loss function, which leverages feature maps from a VGG19 network, helps in achieving results that are not only high in pixel accuracy but also superior in terms of textural and perceptual quality.

In summary, our SRGAN implementation validates the methodology proposed in the referenced paper, showcasing its potential in producing photo-realistic super-resolution images that significantly close the gap between the perceived quality of the upscaled images and their high-resolution targets.

Avg Discriminator Loss	Avg Generator Loss
0.00742	0.8535

VI. Conclusion

In conclusion, our implementation of the Super-Resolution Generative Adversarial Network (SRGAN) has demonstrated remarkable success in generating high-resolution images from low-resolution inputs. Drawing from the pioneering methodologies established by [1] and supplementing them with advancements from subsequent research [2,4,10], we have effectively bridged the gap between the computational pursuit of image super-resolution and the subjective criterion of photo-realism. The perceptual loss function, tailored by insights from the VGG network-based feature maps [13], has been instrumental in transcending traditional pixel-based accuracy, emphasizing the nuanced textures and fine details that contribute to the natural appearance of images. The SRGAN framework has proved to be a robust solution, setting a precedent for future explorations in the domain of image super-resolution and opening avenues for real-world applications that demand high-fidelity image reconstructions. This endeavor has substantiated the potency of GANs in complex image processing tasks and underscored the synergy between deep learning architectures and perceptually-driven loss functions in achieving superior super-resolution outcomes.

VII. Code

Please find link to our project code: here.

ACKNOWLEDGMENT

We wish to extend our deepest gratitude to Professor Rob Fergus, whose invaluable insights and guidance have been instrumental in the fruition of this project.

REFERENCES

[1] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, ´ Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, Wenzhe Shi, "Photo-realistic single image super-resolution using a generative adversarial network" arXiv:1609.04802v5 [cs.CV] 25 May 2017.

[2] Q. Yang, R. Yang, J. Davis, and D. Nister. Spatial-depth super ´ resolution for range images. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1–8, 2007.

[3] W. Zou and P. C. Yuen. Very Low Resolution Face Recognition in Parallel Environment . IEEE Transactions on Image Processing, 21:327–340, 2012.

[4] K. Nasrollahi and T. B. Moeslund. Super-resolution: A comprehensive survey. In Machine Vision and Applications, volume 25, pages 1423–1468. 2014.

[5] C.-Y. Yang, C. Ma, and M.-H. Yang. Single-image super-resolution: A benchmark. In European Conference on Computer Vision (ECCV), pages 372–386. Springer, 2014.

[6] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multi-scale structural similarity for image quality assessment. In IEEE Asilomar Conference on Signals, Systems and Computers, volume 2, pages 9–13, 2003.

[7] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4):600–612, 2004.

[8] P. Gupta, P. Srivastava, S. Bhardwaj, and V. Bhateja. A modified psnr metric based on hvs for quality assessment of color images. In IEEE International Conference on Communication and Industrial Application (ICCIA), pages 1–4, 2011.

[9] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations (ICLR), 2015.

[10] J. Johnson, A. Alahi, and F. Li. Perceptual losses for real-time style transfer and super- resolution. In European Conference on Computer Vision (ECCV), pages 694–711. Springer, 2016.

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

[12] Eirikur Agustsson and Radu Timofte NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study

[13] 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

[14] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In European Conference on Computer Vision (ECCV), pages 184–199. Springer, 2014.

[15] S. Schulter, C. Leistner, and H. Bischof. Fast and accurate image upscaling with super-resolution forests. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3791– 3799, 2015.