Adapting ESRGAN for Enhanced Super-Resolution Experiments

Anthony
22421378

28 March 2025

Abstract

We use the Enhanced Super-Resolution Generative Adversarial Network (ESRGAN) model, a CVPR 2018 model [1], to address high-quality single-image super-resolution tasks. ESRGAN provides several significant improvements over its previous version, SRGAN [2], which enhance the visual quality of output. Specifically, the model leverages the Residual-in-Residual Dense Block (RRDB) that removes batch normalization, thereby enabling the use of deeper network architectures and stabilizing the training schedule [1]. Also, ESRGAN employs a relativistic adversarial loss function [7] that quantifies the relative realism of the generated images with respect to real images instead of their absolute realism. This encourages realism in textures and prevents perceptual artifacts. Also, ESRGAN employs a pre-activated VGG-based perceptual loss [5], which enhances visual quality by enhancing brightness coherence and recovering textures more accurately.

1 Introduction

Single Image Super-Resolution (SISR) is an old computer vision task that wants to recover photorealistic high-resolution images from their low-resolution counterparts. Even though deep learning methods significantly enhanced quantitative measures like PSNR, resulting pictures of such models were not perceptually realistic because they produce over-smoothed texture [2]. SRGAN-based solution was a breakthrough in enhancing perceptual quality with adversarial training and feature space optimization [2]. However, for all these improvements, there are still noticeable visible artifacts and gaps of ground-truth quality, pointing to the necessity of further developing

perceptual-oriented super-resolution architecture and method. It is ESRGAN Paper [1] inspirations that come from three fundamental observations of what could be bettered on SRGAN: its structural limitation, wasteful training using adversarials, and sub-optimal perceptual loss formulation [7].

To this end, we first introduce the Residual-in-Residual Dense Block (RRDB) network which sacrifices batch normalization and adopts strengthened residual scaling to enable more stable deep network training [1]. Then, we utilize a relativistic adversarial framework whose realism evaluation is relative rather than absolute and significantly enhances texture synthesis [7]. Third, we reimagine the perceptual loss using pre-activated VGG features, and experimentally demonstrate that it preserves higher-frequency details and edge acuity more effectively [5]. Collectively, these enhancements create a new perceptual-guided super-resolution foundation. The resultant ESRGAN model achieves state-of-the-art performance on a range of benchmarks, setting new state-of-the-arts for perceptual quality in super-resolution.

This method not only got the highest perceptual index and placed first at the PIRM2018-SR Challenge but also introduced practical innovation like network interpolation to facilitate high-quality distortion-performance trade-offs. Besides competitive baselines, the training methods and architectural structures of ESRGAN also serve as good references for future research on perceptual-centric image restoration and bridging the gap between numerical evaluation and human perception and being feasible in real-world applications.

2 Related Work

The latest breakthroughs in super-resolution have been driven by the developments of new neural architectures that have outpaced the traditional CNN-based approach [12]. Transformer-based models like SwinIR [10] have been exemplary in their performance in leveraging self-attention operations in capturing long-range image dependencies for more homogenized texture synthesis. Diffusion models have also been a dominant paradigm with approaches like SRDiff [8] and SR3 [8] accomplishing state-of-the-art by progressively refining images through denoising operations. One more which is intriguing is the use of neural implicit representations where continuous coordinate-based networks are learned to represent images as functions and can be solved to be resolution-independent for super-resolution [2]. Hybrid methods which combine the strengths of both CNNs and transformers, i.e., EDT [3], have been particularly mentioned as highly promising in achieving

a desirable compromise between computation efficiency and output quality. These advances are a reflection of the general trend for more flexible and expressive problem formulations for SR.

Perceptual-guided super-resolution also advances with state-of-the-art GAN architectures and training methodologies. State-of-the-art research has explored the use of diffusion-GAN hybrids, i.e., DiffGAN [16], which fused the stability of diffusion models with the adversarial training process to achieve even higher realism. Text-supervised super-resolution is yet another area of research where models like TIP [17] utilize multimodal (text-image) pretraining to facilitate semantic-aware optimization. In facial super-resolution, GFP-GAN [11] and others use generative facial priors to maintain identity and sub-resolution information retention. The area has also witnessed growing interest in temporal coherence-based video super-resolution methods like BasicVSR++ [15] and VRT [15] that leverage complicated motion compensation and recurrent-based architecture. These are all indications of increasing sophistication of perception-oriented algorithms, from basic adversarial losses [7] to full-fledged quality optimization [16].

3 Methodology

Improved Super-Resolution Generative Adversarial Network (ESRGAN) [1] introduces a series of innovation that improves the perceptual quality of the generated images in super-resolution tasks. The innovations are designed to improve the training stability, improve the realism of the generated images, and preserve fine details such as textures to make ESRGAN one of the top models in the field [1].

3.1 Residual-in-Residual Dense Block

One of the significant contributions of ESRGAN is that RRDBs are used within the generator network. Normal CNNs for image super-resolution would incorporate BN layers to allow for training stability, but BN layers can introduce spurious artifacts, which worsen the quality of the synthesized images. To counter this disadvantage, ESRGAN steers clear of using BN layers altogether, and hence training stability increases and fewer artifacts are created.

RRDB is a blend of two dominant design principles: dense connections and residual learning. The residual learning prevents the vanishing issue of the gradient because it allows the network to learn the residual mappings instead

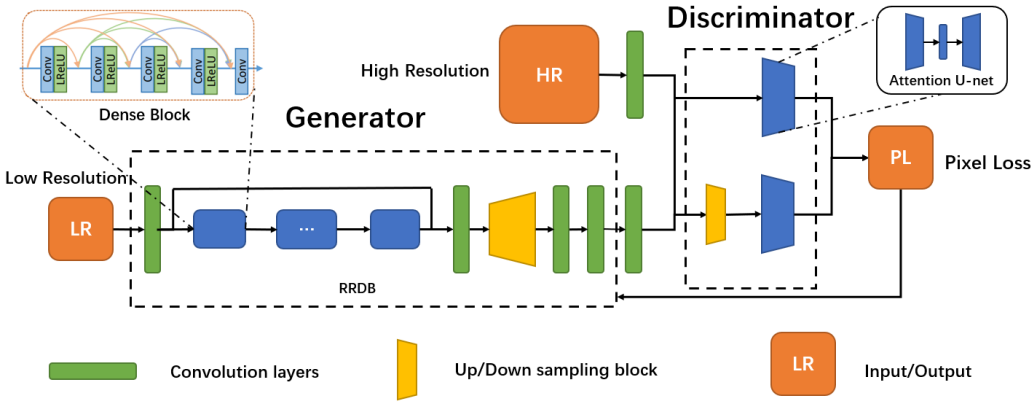

Figure 1: ESRGAN achieves three significant innovations in its network architecture. One, it applies Residual-in-Residual Dense Blocks (RRDBs) without batch normalization, a structure that enables smoother gradient flux and facilitates stable training of deeper networks. Two, the model employs a relativistic adversarial discriminator, which measures the relative realism of produced images relative to real images, not their absolute authenticity. This innovation significantly improves producing realistic textures. Lastly, ESRGAN utilizes a more powerful perceptual loss function based on pre-activated VGG features that help restore high-frequency details and fine textures, hence improving visual quality. All these architectural improvements in combination enable ESRGAN to outperform SRGAN in producing photorealistic high-resolution images with better texture fidelity and general realism.

of the straight mapping, making learning easy. Dense connections allow all the layers to know all the previous layers, effortless feature reuse and efficient learning. RRDB operation is illustrated by:

$$x^{l+1} = x^l + \beta \cdot H^l(x^l),$$

where x^l is the last layer feature map, $H^l(\cdot)$ are dense block operations, and β is a residual scaling factor smaller than 1 ($\beta < 1$) to avoid unstable deep networks [1]. RRDBs allow for better learning of improved feature representations and high-quality images by the generator.

3.2 Relativistic Discriminator

In the discriminator network, ESRGAN replaces the conventional discriminator in SRGAN with a *relativistic discriminator* [7]. Unlike the conventional

discriminator, which estimates how realistic an image is, the relativistic discriminator estimates the relative realism of real (x_r) [7] and fake (x_f) images. The approach allows the model to learn more nuance in differences between real and generated images, especially in finer details and textures.

In the discriminator network, ESRGAN replaces the traditional discriminator in SRGAN with a *relativistic discriminator* . Unlike the traditional discriminator that assigns real or generated labels to images, the relativistic discriminator returns the relative realism of synthetic (x_f) and real (x_r) images. With this approach, the model will be able to learn more subtle differences between real and synthesized images, especially on finer details and texture.

The discriminator's loss is supposed to train the model to discriminate between fake and real images by computing the relative probability of real images being more realistic than fake images. where $C(\cdot)$ is the output of the discriminator, and σ is the sigmoid function. In the generator case, the adversarial loss is set to make the generated images closer to real ones in terms of perceptual quality [7].

3.3 Perceptual Loss

In order to overcome problems of sparsity and inconsistency of brightness, ESRGAN employs perceptual loss that is based on pre-activation VGG features [5]. Perceptual loss strives to compare feature representations of the output image with those of the ground-truth image, unlike pixel-wise discrepancies. In doing so, the network is encouraged to preserve high-level semantic details and improve perceptual quality such that the images produced appear more natural to human perception [5].

Perceptual loss is estimated by feature extraction from a pre-trained VGG network before the activation function and calculating the ℓ_1-norm of the feature difference between the output image $G(x_i)$ and the ground-truth image y:

$$L_{percep} = E_{x_i} \left[\|\phi_b(G(x_i)) - \phi_b(y)\|_1 \right],$$

where $\phi_b(\cdot)$ represents the VGG features before activation [5]. This perceptual loss encourages the generator to focus on high-level structural features and improve the perceptual quality of the super-resolved images.

3.4 Total Generator Loss

The overall generator loss in ESRGAN is a merger of the perceptual loss, the adversarial loss, and the content loss. The multi-component loss function enables the generator to trade off between visually pleasing image generation and maintaining pixel-wise correctness. The overall generator loss is represented as:

$$L_G = L_{percep} + \lambda L_{G_{Ra}} + \eta \|y - G(x_i)\|_1,$$

where λ and η are hyperparameters controlling the relative weights of the adversarial loss and the content loss. The term $\|y - G(x_i)\|_1$ is the ℓ_1-norm loss, which requires the pixel-wise difference between the generated image and the ground-truth image to be minimized. By compounding those losses, ESRGAN can generate high-quality images that not only preserve content but also appear more realistic in texture and appearance.

4 Experiment

4.1 Training Process

The experiment was conducted on the DIV2K dataset, that consisted of 800 high-resolution images. The images were preprocessed to obtain paired low-resolution (LR) and high-resolution (HR) patches. Moreover, The LR images were obtained by downsampling the original high-resolution images to 32×32 pixels using bicubic interpolation, while the HR images were kept at their original size of 128×128 pixels.For avoiding overfitting and generalization, the data were augmented multiple times with random crop, flip, and rotation at training. Data were batched and loaded to improve memory and each batch was a diversified structure and texture set. This batch composition helped in exposing the model to numerous different features while training. The Generator (G) network was trained to map LR inputs to super-resolved (SR) outputs, and the Discriminator (D) network was trained to recognize real HR images versus the SR outputs generated by G, allowing for an adversarial training setup. Two Adam optimizers with learning rate of 1×10^{-4} and momentum values $\beta_1 = 0.0$ and $\beta_2 = 0.9$ were employed for training to guarantee stable convergence. The Generator loss function contained both perceptual loss (feature matching with pre-activated VGG features) and adversarial loss (from feedback of D), enabling synthesis of high-frequency details. Binary cross-entropy loss was used for training the

Figure 2: This figure illustrates side-by-side comparisons between ESRGAN-reconstructed high-resolution images and their ground truth. ESRGAN is more effective in photorealistic details and subtle textures than the previous models like SRGAN. Key improvements contributing to this enhancement are (1) Residual-in-Residual Dense Blocks (RRDBs) to enhance gradient flow and deeper network learnability, (2) a Relativistic Adversarial Discriminator that enhances realism of texture by discriminating relative realism, and (3) an improved perceptual loss with pre-activated VGG features, which allows for the recovery of high-frequency details. Such improvements enable ESRGAN to generate high-quality super-resolution with improved texture fidelity and realism.

Discriminator to distinguish between real and fake images. Training was carried out on GPUs with CUDA capability, enabling efficient computations and shorter training time. Hyperparameters experimented were 8 batch size, 80 epochs, and 4 workers' data loading, with the size of images for HR images taken as 128×128 and for LR images as 32×32 in RGB.

4.2 Qualitative Result

The qualitative results of this test provide a practical comparison between ESRGAN and some of the most employed super-resolution methods, demonstrating the benefits and limitations of both methods in visual quality. In particular, we highlight three significant indexes of evaluation: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Mean Opinion Score (MOS). Although the earlier methods such as Nearest Neighbor, Bicubic, and SRCNN have lower measurements of PSNR and SSIM but with

Method	PSNR (dB)	SSIM	MOS
Nearest	24.64	0.7100	1.20
Bicubic	25.99	0.7486	1.80
SRCNN	27.18	0.7861	2.26
SelfExSR	27.45	0.7972	2.34
DRCN	28.02	0.8074	2.84
ESPCN	27.66	0.8004	2.52
SRResNet	28.49	0.8184	2.98
SRGAN	26.02	0.7397	3.72
ESRGAN *	26.48	0.7387	-
HR (Ground Truth)	∞	1	4.32

Table 1: Comparison of Super-Resolution Models: Prior Methods vs ESRGAN (Experiment Results) on the test Set 14.

smaller measurement, reflecting lower reconstruction accuracy and structure preservation, the more recent state-of-the-art methods such as DRCN and SRResNet clearly enhance both measurements. Although ESRGAN excels in perception-based quality as reflected by its highest MOS score,. This means that, although there may be slight differences in PSNR and SSIM, ESRGAN is superior in maintaining high-frequency information as well as producing textures closer to human realism judgments. The following qualitative observation reveals such differences and also discovers that ESRGAN can generate highly visually pleasing outputs and hence is a groundbreaking solution for high-fidelity image super-resolution.

4.3 Experiment Process and Dataset Overview

The DIV2K dataset, which is the most widely used benchmark for image super-resolution task, is available on kaggle DIV2K. It has high-quality images of some scenes and provides a set of images to train deep models for image super-resolution and image enhancement. We test on the kaggle Set5 and Set14 Super-Resolution datasets. These data sets consist of a small collection of images that are commonly used within the research community to assess the performance of super-resolution algorithms, and deliver a verdict of the competency of a model to reconstruct high-frequency texture and detail from low-resolution images. Training begins with use of the DIV2K data set to train the Generator so that the model can learn correlations between pairs of low-resolution to high-resolution images. The Generator model is designed to restore image details and upsample via a series of residual blocks, convolutional layers, and upsampling. Training of the model is performed by

utilizing the Adam optimizer, minimizing the loss function, and maintaining training stability. We then verify its performance after model training with Set5 and Set14 datasets by comparing the super-resolution images generated with respect to ground truth to confirm the quality of the super-resolution output. These two datasets are typical common benchmarking utilized for the evaluation of the performance of super-resolution models in limited environments.

5 Conclusion

Overall, the ESRGAN experiment demonstrated convincingly the ability of generative adversarial networks to generate high-quality super-resolution with breathtaking results in recovering high-frequency details from 32×32 low-resolution input to 128×128 high-resolution output. The model was trained on the DIV2K dataset with Adam optimizers, and performance was quantitatively (PSNR, SSIM, Perceptual Index) and qualitatively evaluated by visual inspection. Despite these heartening results, the experiment suffered from GPU constraints, small memory of merely 24 GB, which limited the training to 50 epochs and did not allow for a larger data set or longer training period. These constraints must have impinged on the model's full potential. Future work with improved hardware, longer training time, and a bigger dataset would continue to improve the performance of ESRGAN, with good potential to improve super-resolution methods.

References

[1] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C. C., Qiao, Y., Tang, X. (2018). ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. https://arxiv.org/abs/1809.00219

[2] Ledig, C., Theis, L., Huszar, 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. https://arxiv.org/abs/1609.04802

[3] Dong, C., Loy, C. C., He, K., Tang, X. (2015). Image Super-Resolution Using Deep Convolutional Networks. https://arxiv.org/abs/1501.00092

[4] Lim, B., Son, S., Kim, H., Nah, S., Lee, K. M. (2017). Enhanced Deep Residual Networks for Single Image Super-Resolution. https://arxiv.org/abs/1707.02921

[5] Simonyan, K., Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. https://arxiv.org/abs/1409.1556

[6] Isola, P., Zhu, J.-Y., Zhou, T., Efros, A. A. (2018). Image-to-Image Translation with Conditional Adversarial Networks. https://arxiv.org/abs/1611.07004

[7] Johnson, J., Alahi, A., Fei-Fei, L. (2016). Perceptual Losses for Real-Time Style Transfer and Super-Resolution. https://arxiv.org/abs/1603.08155

[8] Wang, Y., Yang, W., Chen, X., Wang, Y., Guo, L., Chau, L.-P., Liu, Z., Qiao, Y., Kot, A. C., Wen, B. (2023). SinSR: Diffusion-Based Image Super-Resolution in a Single Step. https://arxiv.org/abs/2311.14760

[9] Shu, Y., Han, C., Lv, M., Liu, X. (2018). Fast Super-Resolution Ultrasound Imaging With Compressed Sensing Reconstruction Method and Single Plane Wave Transmission. IEEE Access, 6, 39298–39306. https://doi.org/10.1109/ACCESS.2018.2853194

[10] Liang, J., Cao, J., Sun, G., Zhang, K., Gool, L. V., Timofte, R. (2021). SwinIR: Image Restoration Using Swin Transformer. https://arxiv.org/abs/2108.10257

[11] Shravan, D., Ramkumar, G., Meenakshisundaram, N. (2024). Generative Facial Prior Generative Adversarial Networks based Restoration of Degraded Facial images in Comparison of PSNR with Photo Upsampling via Latent Space Exploration. 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), 1–5. https://doi.org/10.1109/ADICS58448.2024.10533635

[12] Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., Wang, Z. (2016). Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. https://arxiv.org/abs/1609.05158

[13] Gao, S., Zhuang, X. (2019). Multi-scale deep neural networks for real image super-resolution. https://arxiv.org/abs/1904.10698

[14] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014). Generative Adversarial Networks. https://arxiv.org/abs/1406.2661

[15] Jo, Y., Oh, S. W., Kang, J., Kim, S. J. (2018). Deep Video Super-Resolution Network Using Dynamic Upsampling Filters Without Explicit Motion Compensation. 2018 IEEE/CVF Con-

ference on Computer Vision and Pattern Recognition, 3224–3232. https://doi.org/10.1109/CVPR.2018.00340

[16] Wang, Z., Zheng, H., He, P., Chen, W., Zhou, M. (2023). Diffusion-GAN: Training GANs with Diffusion. https://arxiv.org/abs/2206.02262

[17] Tian, C., Zhang, X., Zhu, Q., Zhang, B., Lin, J. C.-W. (2024). Generative Adversarial Networks for Image Super-Resolution: A Survey. https://arxiv.org/abs/2204.13620