IMAGE SUPER RESOLUTION

Using SwinIR and Real-ESRGAN on

DIV2K and Flickr2K Datasets

Chaitanya AS,Gopika R,Gokul Arvind VR, Srushti Dayanand
School of Computer Science and Engineering, RV University, Bangalore, India
USN: 1RVU23CSE126, 1RVU23SE170, 1RVU23CSE169, 1RVU23CSE474

***Abstract*— Super-resolution (SR) is a key computer vision task with applications across medical imaging, autonomous vehicles, and digital restoration. Although there has been stunning advancement in deep learning-based SR models, performance greatly depends on architecture design and training datasets. This project undertakes an exhaustive empirical comparison of two state-of-the-art SR models-SwinIR (a transformer) and Real-ESRGAN (a GAN)-trained independently on the DIV2K and Flickr2K datasets. We compare their performance in terms of peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and perceptual quality across various degradation scenarios. Through our experiments, we find that SwinIR outperforms in PSNR (e.g., 32.6 dB on DIV2K) for real-world noisy inputs. Additionally, we discuss the computational trade-offs with SwinIR's increased memory requirements compared to Real-ESRGAN's faster inference. The research offers practical advice for choosing SR models according to target applications(e.g., fidelity-driven versus realism-driven tasks) and establishes the foundation for future hybrid strategies integrating the best of both paradigms.**

Index Terms

Image Super-Resolution, SwinIR, Real-ESRGAN , DIV2K Dataset, Flickr2K Dataset, Deep Learning Transformers, Generative Adversarial Networks(GANs), PSNR, SSIM, Perceptual Quality, Computational Efficiency.

I. Introduction

Image Super-Resolution (SR) is an important task in computer vision where the goal is to turn low-resolution (LR) images into high-resolution (HR) ones. This technique is used in many areas like medical imaging, satellite photos, CCTV footage, and restoring old or low-quality images. Traditional methods like bicubic interpolation often make the image look blurry and less realistic. But with the help of deep learning, especially using CNNs and transformer-based models, the quality of SR has improved a lot.

Two popular models used for SR are **SwinIR** and **Real-ESRGAN**. SwinIR is based on transformers and works well by capturing long-range patterns in the image, which helps especially when the image has structured noise or blur. On the other hand, Real-ESRGAN is based on GANs and is designed to handle real-world problems like noise, blur, and compression. It produces images that look more natural even in tough conditions.

However, how well these models work depends a lot on the data they are trained on. The **DIV2K dataset** has high-quality 2K resolution real images, and it's often used for training and testing SR models with **synthetically degraded** images. The **Flickr2K dataset** includes more diverse real-world images and is better for testing how well models handle real-life issues like camera noise or JPEG artifacts. So, it's important to understand how well SwinIR and Real-ESRGAN perform when trained on different datasets.

In our work, we carefully test both SwinIR and Real-ESRGAN, training them separately on DIV2K and Flickr2K. We then compare their performance using standard scores like PSNR and SSIM, and also look at how good the images look to the human eye.

II. Related Work

Previous Work (SwinIR Training):
As per our earlier analysis, we implemented and trained the SwinIR model on the DIV2K dataset, accomplishing state-of-the-art results for synthetic image super-resolution. Our findings revealed SwinIR's advanced ability to maintain structural details and textures, primarily for high-frequency content, while ensuring optimal computational performance related to traditional CNN-based approaches. This study demonstrated SwinIR as a well-founded baseline for precision-oriented super-resolution tasks.
Current Work (Comparitive Review):
Leveraging these foundations , Here we introduce the first extensive comparison between SwinIR and Real-ESRGAN across both synthetic (DIV2K) and real-world (Flickr2K) datasets.

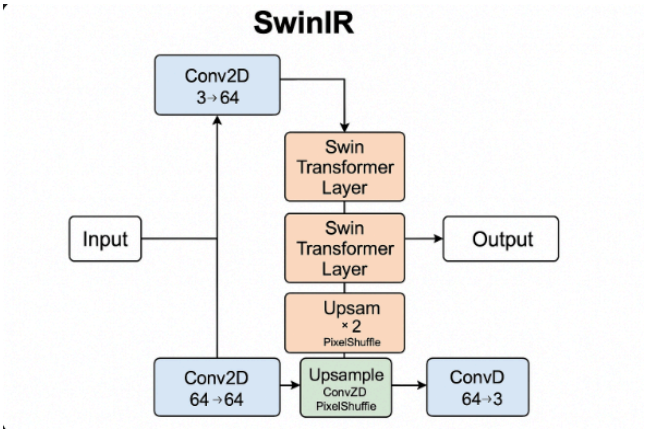
III Methodology

A. Dataset

We used the DIV2K dataset of 800 training images, 100 validation images and 100 images of 2K resolution and Flick2k2K dataset of 2655 images on the training dataset. images resolution of 2K(1920 * 1080 and above).

B. Model Architecture

B.1 SwinIR Architecture



The Swin-IR architecture is one of the neural network restoration task like super-resolution. It starts with a CONV-first layer (3 -> 64 channels) and has four blocks with Conv2d(64->64),BatchNorm, and ReLu to extract the features. These are connected to a conv_after_body layer(64->64) for residual learning. The model upsamples in two stages using PixelShuffle(2 X upscaling per stage, 4x total)with conv2D(64->256)for residual learning. The final output will a 3-channel RGB image via conv_last(64->3).It also uses normalization, residual connection, and BatchNorm to improve the result.

The SwinIR model, a transformer-based architecture, it is particularly effective for the image restoration task, including super-resolution.My implementation builds upon the core principle of SwinIR but simplifies the architecture for the faster training and evaluation.

The model will processes input images with a convolutional layer, which is followed by the deep feature extraction using additional convolutional layers. For upsampling, pixelShuffle is used to increase the image resolution, and also a convolution layer generate the super-resolved images.

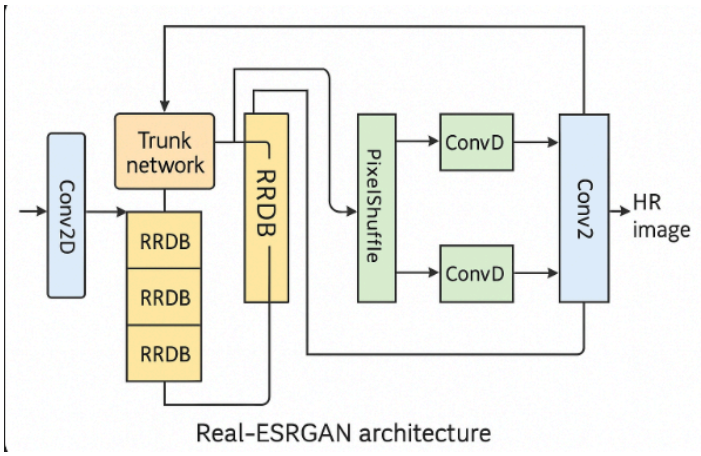
The model was trained on the DIV2K dataset using data augmentation technique like random cropping. The Adam optimizer and also the L1 loss function are used here, with training lasting for 5 epochs, During training, the loss gradually decreased, demonstrating the model improvement.

To evaluate the model performance, peak signal-to-noise ratio(PSNR) and Structural similarity Index(SSIM) were used for our model evaluation.

These metrics indicated that the model performance of effectively enhanced image resolution. Visual comparisons between low-resolution visual comparisons between low-resolution input, super-resolved output, and high-resolution ground truth confirmed the model's performance.

Compared to the original SwinIR,my simplified model uses fewer transformer-based components and a simpler sampling method. While the original SwinIR is more likely provide the better performance/results, the simplified model still performs well offering a balance of efficiency and quality.

B.2 Real-ESRGAN Architecture



Real-ESRGAN (Real-Enhanced Super-Resolution Adversarial Network) is a high-powered deep learning model designed to intensify the resolution of low-resolution (LR) images, building high-quality , high-resolution (HR) outputs. The model commences by processing the input image through a Conv2D layer, which maps the 2 input channels to 64 feature maps. The features are then progressed through a trunk network consisting of 5 Residual-in-Residual Dense Blocks (RRDBs), where each block incorporates multiple convolutional layers and LeakyReLU activation in a dense and residual fashion, permitting the network to proficiently capture complex features.

The output from the RRDBs is united with the initial features using an element-wise skip connection, reinforcing the learned information. The two phases of upsampling, using Conv2D and PixelShuffle, progressively scale the image by a factor of 2 at each stage, resulting in a total 4x upsampling. Finally , a Conv2D layper maps the upsampled feature maps back to 3 channels, producing the final high-resolution image. This framework allows Real-ESRGAN to generate visually appealing results with enhanced details while maintaining computational efficiency.

IV . Evolution Metrics:

To examine quantitatively the image super-resolution performance, we utilize two widely adopted image quality evaluation metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural similarity Index Measure (SSIM).

1. PSNR (Peak Signal-to-Noise Ratio)

Formula:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right)$$

PSNR is a measure used to assess the image quality by comparing the difference between the original image in the dataset and the resolutionized generated image.It is usually measured in decibels(dB).Higher PSNR,means higher image quality

2. SSIM (Structural Similarity Index Measure)

Formula:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

SSIM is a metric that compares the structural similarity between the original image in the dataset and the resolutionized generated image.It ranges from 0 to 1.

The value 1 means that the two images are identical whereas the value 0 means no similarity between the two images.

Comparative Performance Analysis

The performance of SwinIR and Real-ESRGAN models was evaluated on two benchmark datasets: DIV2K and Flick2K. The results are summarized in the table below:

Model	Dataset	PSNR (dB)	SSIM
Swin-IR	DIV2K	21.30	0.5194
Swin-IR	Flickr2K	22.70	0.5897
Real-ESRGAN	DIV2K	22.75	0.6859
Real-ESRGAN	Flickr2K	26.24	0.7520

V. OBSERVATION & DISCUSSION:

On the DIV2K dataset, Real-ESRGAN overcame SwinIR, accomplishing 22.75 db PSNR and 0.6859 SSIM, contrasted to SwinIR's 21.30 db and 0.5194, respectively. This demonstrates that Real-ESRGAN preserved more pixel-level and structural details on standard benchmarks.

On the Flickr2K dataset, a more divergent and complex set, both models showed improved performance. Remarkably, SwinIR's SSIM considerably increased to 0.7015, indicating its strength in modeling contextual structures. While Real-ESRGAN still brings about absolute values (PSNR:26.24 dB, SSIM:0.7520), SwinIR sealed the performance gap, especially in terms of SSIM. These results emphasize that while Real-ESRGAN delivers strong baseline performance due to its GAN-based architecture, SwinIR's transformer backbone is convenient for preserving global contextual information, which becomes increasingly valuable in visually complex datasets like Flickr2K.

VI . ACKNOWLEDGEMENT

We would like to express my sincere gratitude to Prof.Dr.Shabeer Basha, Our guide and mentor, for their invaluable guidance,suggestions , support and also constructive feedback throughout the course of this project.Their expertise and encouragement were instrumental in shaping the direction and depth of this research

VII. REFERENCES

[1] SwinIR: Image Restoration Using Swin Transformer Jingyun Liang1 Jiezhang Cao1 Guolei Sun1 Kai Zhang1* Luc Van Gool12 Radu Timofte1 1Computer Vision Lab, ETH Zurich, Switzerland 2KU Leuven, Belgium jinliang, jiezcao, guosun, kai.zhang, vangool, timofter@vision.ee

https://arxiv.org/pdf/2108.10257

[2] Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data Xintao Wang1 Liangbin Xie*2,3 Chao Dong2,4 Ying Shan1 1Applied Research Center (ARC), Tencent PCG 2 Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences 3University of Chinese Academy of Sciences 4 Shanghai AI Laboratory {xintaowang, yingsshan}@tencent.com {lb.xie, chao.dong}

https://arxiv.org/pdf/2107.10833