VIDEO SUPER-RESOLUTION USING A GENERATIVE ADVERSARIAL NETWORK

Mahendra Pratap Singh (24m1226), Saad Bin Ayaz (24m1067), Nirmal S (20d070057), Joel Anto Paul (210070037), Ashwin Prajapati (24m1064)

ABSTRACT

Despite advances in speed and accuracy of single image super-resolution using deep convolutional neural networks, recovering fine texture details at large upscaling factors remains a challenge. Traditional methods focus on minimizing mean squared error, resulting in high PSNR but perceptually unsatisfying images that lack high-frequency detail. To address this, SRGAN—a generative adversarial network for 2,4,8× image super-resolution—introduces a perceptual loss combining adversarial and content losses. The adversarial loss encourages outputs that lie on the natural image manifold, while the content loss focuses on perceptual rather than pixel-wise similarity. SRGAN's deep residual network successfully recovers realistic textures, and extensive mean-opinion-score tests show it significantly outperforms existing methods in perceptual quality.

1. INTRODUCTION

Super-resolution (SR) refers to the process of estimating a high-resolution (HR) image from its low-resolution (LR) counterpart, a task that remains highly challenging, especially for large upscaling factors. While deep convolutional neural networks (CNNs) have significantly advanced SR in terms of accuracy and speed, most approaches optimize for pixel-wise loss functions such as mean squared error (MSE), which leads to high peak signal-to-noise ratio (PSNR) but often fails to recover fine texture details, resulting in perceptually unsatisfying images. To address this, we explore the use of generative adversarial networks (GANs) for SR and introduce SRGAN—an architecture that combines a deep residual network with a novel perceptual loss function. This loss includes both a content loss based on high-level feature representations from a pre-trained VGG-16 network and an adversarial loss that encourages the output to reside on the manifold of natural images. SRGAN is the first framework capable of producing photo-realistic images for 2,4,8× super-resolution, demonstrating superior perceptual quality as validated by extensive Mean Opinion Score (MOS) tests across standard benchmark datasets.

Recent advances in super-resolution (SR) have shifted from CNN-based models to Transformer-based approaches due to their success in modeling long-range dependencies. However, studies show that existing Transformer models like SwinIR do not fully exploit input context and may suffer from artifacts. To address these limitations, the Hybrid Attention Transformer (HAT) is proposed. HAT combines self-attention with channel attention and introduces an overlapping cross-attention module to enhance local and global feature interactions. Furthermore, a same-task large-scale pre-training strategy is adopted to better unlock the potential of Transformers for SR. These innovations enable HAT to achieve state-of-the-art results, significantly surpassing previous models by 0.3–1.2 dB in PSNR.

2. PROBLEM STATEMENT

To develop and compare two distinct models—SRGAN and a Transformer-based model—for enhancing the resolution of low-quality videos up to 2K (2048x1080), ensuring both structural and perceptual quality are preserved or enhanced. This project explores the enhancement of video resolution from low-quality inputs to 2K resolution using two deep learning approaches: SRGAN, a model based on a generative adaptive network, and a model based on a transformer, specifically the hybrid attention transformer (HAT). The goal is to assess the reconstruction quality, visual fidelity, and performance trade-offs of both approaches on standard benchmark datasets. Evaluation metrics include PSNR, SSIM, and perceptual quality assessments.

3. METHODOLOGY

3.1. SRGAN Approach

SRGAN, a Generative Adversarial Network (GAN) for super-resolution that generates photo-realistic images — much more visually pleasing than those optimized only for MSE (Mean Squared Error) or PSNR. SRGAN is a GAN-based architecture where the generator is initialized with SRResNet weights, then it is fine tuned using perceptual loss. Perceptual loss contains two types of loss, i.e. content loss and adversarial loss. Content loss is calculated based on differences in the VGG16 feature maps rather than pixel-wise MSE. Adversarial loss uses a discriminator to push generated images towards the natural image manifold.

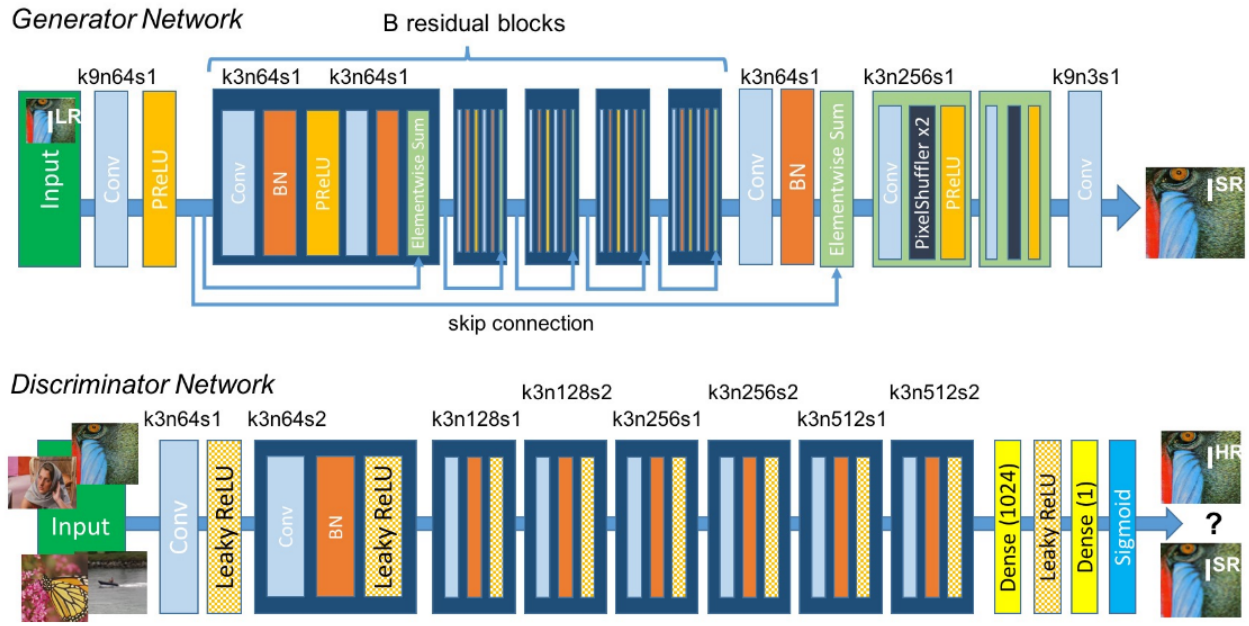

Fig. 1: Architecture of Generator and Discriminator Network

3.2. SRGAN Architecture

The SRGAN architecture has three different components:

*3.2.1. **Adversarial network architecture***

In this approach, a discriminator network which we optimize in an alternating manner along with generator network to solve the adversarial min-max problem. The general idea behind this formulation is that it allows one to train a generative model with the goal of fooling a differentiable discriminator that is trained to distinguish super-resolved images from real images. With this approach our generator can learn to create solutions that are highly similar to real images and thus difficult to classify by Discriminator. This encourages perceptually superior solutions residing in the subspace, the manifold, of natural images. This is in contrast to SR solutions obtained by minimizing pixel-wise error measurements, such as the MSE.

At the core of our very deep generator network G, which is illustrated in Figure 1 are B residual blocks 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. We increase the resolution of the input image with two trained sub-pixel convolution layers.

To discriminate real HR images from generated SR samples we train a discriminator network. The architecture is shown in Figure 1. The discriminator network is trained to solve the maximization problem. 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 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.

*3.2.2. **Perceptual Loss function***

The definition of our perceptual loss function is critical to the performance of our generator network. While perceptual loss is commonly modeled based on the MSE, it design a loss function that assesses a solution with respect to perceptually relevant characteristics. We formulate perceptual loss as the weighted sum of content loss and adversarial loss.

3.3. Transformer Approach (HAT)

The Hybrid attention Transformer combines different types of attention mechanism for image super-resolution. It combines channel and self-attention for richer feature representation. It uses overlapping cross-attention to improve window interaction.

It consists of many important blocks:

3.3.1. *Residual Hybrid Attention Group (RHAG)*

Each RHAG serves as a modular unit in the deep feature extraction stage of HAT, containing multiple Hybrid Attention Blocks (HABs) and one Overlapping Cross-Attention Block (OCAB), followed by a convolution layer with a residual connection to enhance information flow and representation capacity. By integrating HABs (which fuse window-based self-attention with channel attention) and OCAB (which strengthens cross-window communication), RHAGs enable the network to capture both fine-grained details and global consistency—crucial for super-resolution tasks. Each RHAG includes a residual connection from its input to output, which stabilizes gradient flow and facilitates deeper model stacking without degradation, improving both convergence and final performance.

3.3.2. *Hybrid Attention Block (HAB)*

The HAB integrates a Channel Attention Block (CAB) with the standard Swin Transformer architecture. The CAB is placed after the first LayerNorm layer and works in parallel with the window-based multi-head self-attention (W-MSA) module. This addition helps the network focus on more important channels, thereby enhancing its representation power. The HAB employs a self-attention mechanism inside local windows, allowing the model to focus on small regions of the image. It uses a shifted window approach to enhance the connections between neighboring windows, which increases the range of pixels considered for attention. This strategy significantly improves the model's ability to capture global context and fine details in the image.

3.3.3. *Overlapping Cross Attention Block (OCAB)*

OCAB consists of an Overlapping Cross-Attention (OCA) layer and a Multi-Layer Perceptron (MLP) layer. The OCA layer is designed to establish cross-window connections and enhance the window self-attention mechanism. The MLP layer follows the OCA layer and contributes to the transformation and integration of features. OCA partitions the input feature X into non-overlapping windows for queries (XQ) and overlapping windows for keys (XK) and values (XV). The partitioning is controlled by specific window sizes and a constant to determine the size of overlapping windows relative to the query windows. OCA computes key/value pairs from a larger field compared to the query, facilitating cross-attention operations within each window feature.

4. IMPLEMENTATION APPROACH

4.1. Datasets

At first, We are using DIV2K Image data set to train our model. For calculating the content loss, model is using VGG-16 feature map. We are taking youtube 360p videos for testing purpose. Later for optimizing the performance of our model, we have divided the different videos in different classes. We are taking a data set of YouTube cricket and YouTube video lectures of 1080p resolution.

4.2. Implementaion

4.2.1. *SRGAN Approach*

Our focus was only to enhance quality of 360p videos of YouTube. So for modelling the downscaling by youtube on high resolution video, we are implementing Gaussian filter(sigma=2.5) followed by Bicubic or Lanczos filtering. We are here testing the effect of Bicubic and Lanczos filter over super resolution. Since this implementation takes into account of any kind of youtube 360p videos which may be the reason of latency we are getting in testing videos. We are now dividing videos into different classes of videos. We are now training our model to two specific class of videos which later can be generalized to any class videos. The idea here is to reduce the number of layers the model is using since it is for general case. At input level , we can implement interface which will activate the model for different classes of videos by choosing their feature map.

4.2.2. *Transformer Approach*

- **Dataset Preparation** To build a robust training dataset for super-resolution, a collection of lecture videos in both 1080p and 360p resolutions was sourced from YouTube. These videos provided a diverse range of content, ensuring variability in lighting, text clarity, and visual features commonly found in educational materials. Individual frames were extracted from the videos to create a comprehensive dataset of high-resolution images. This frame-level extraction helped preserve temporal and visual details crucial for training a high-performance super-resolution model.

- **Training Pipeline** The training process began by resizing the low-resolution 360p frames to a spatial dimension of 64×64 pixels, while the corresponding high-resolution 1080p frames were resized to 256×256 pixels. This resizing step facilitated a 4× super-resolution setup, where the model learns to map low-resolution inputs to their high-resolution counterparts. The model was trained using the Adam optimizer, which is well-suited for image restoration tasks due to its adaptive learning rate mechanism. Huber loss was employed as the loss function, offering a balance between mean squared error and mean absolute error, thereby providing robustness against outliers while ensuring smooth convergence. The network was trained to reconstruct

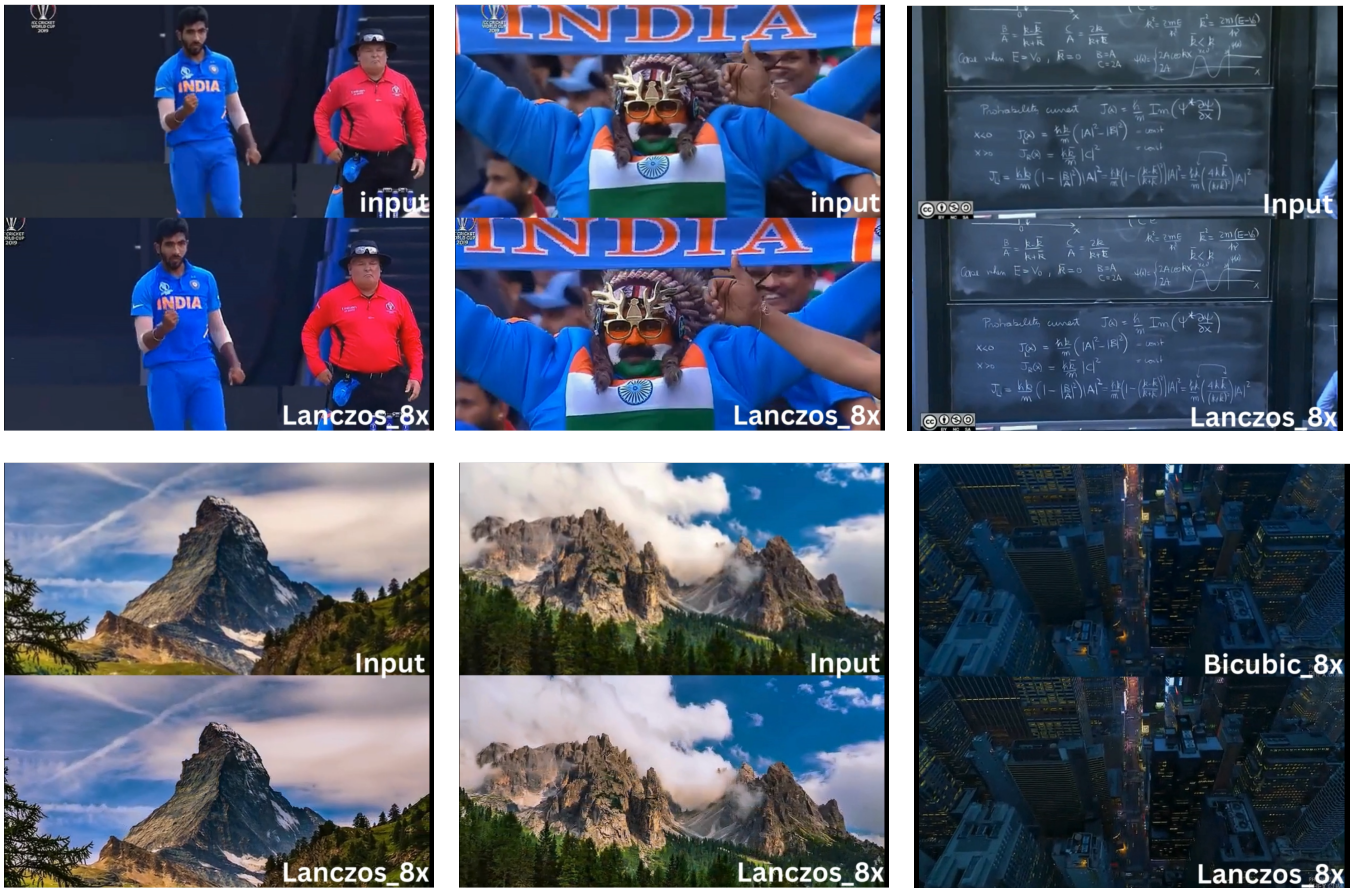

Fig. 2: Results using Bicubic and Lanczos interpolation

detailed high-resolution outputs from heavily downscaled inputs.

- **Testing Phase** During the evaluation phase, the trained model was tested on previously unseen low-resolution videos to assess its generalization performance. The model processed each video frame individually, generating high-resolution reconstructions through frame-by-frame inference. Once all frames were upscaled, they were sequentially combined to regenerate the complete video in high resolution. This reconstruction approach ensured that the temporal consistency and visual fidelity of the videos were preserved in the final output.

5. RESULTS

We have included the Results using Bicubic and Lanczos interpolation filter. Results are shown in Fig.2.

6. CONCLUSION

We have trained the Model using SRGAN approach. We have found that SRGAN Model takes care of every class of video but it takes time to generate the video output. We found that if we divide the class of videos in multiple classes then by implementing the class interface we can reduce the feature map size which will in turn reduce the layer of Models. This will result in reduction of time taken by model to generate the output video.

7. REFERENCES

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