MINOR FINAL PROJECT REPORT SUMMARY

15B19CI691

Supervisor : **Aparajita Nanda**

DCSRGAN

Dimensionally Invariant Color Bounded Super-Resolution Generative Adversarial Networks

GROUP G117 - RAJAT GOEL 18103108 B3, SARTHAK SHARMA 18103112 B3, SHREYA GROVER 18103113 B3

Deployed at github.com/grajat90/ResampleGAN

INTRODUCTION

Image resampling is the process of altering an image's resolution by either adding or removing extra pixels. Increasing the number of pixels in an image is often referred to as upsampling, reducing the number of pixels is referred to as downsampling

Image resampling finds uses in many fields. In computer graphics, it allows texture to be applied to surfaces in a CG imagery, without the need to explicitly model the texture. In medical and remotely sensed imagery it allows an image to be registered with some standard coordinate system.This is primarily done to organize for further processing.

The highly challenging task of estimating a high- resolution (HR) image from its low-resolution (LR) counterpart is mentioned as super-resolution (SR). SR received substantial attention from within the pc vision research community and features a wide selection of applications.

In this work, we propose Dimensionally Invariant Colour Bounded Super-Resolution Generative Adversarial Networks (DCSRGAN) that we employ a deep residual network. Our deep residual network is in a position to recover photo-realistic textures from heavily downsampled images on public benchmarks. The adversarial loss pushes our solution to the natural image manifold employing a discriminator network that's trained to differentiate between the super-resolved images and original photo-realistic images. We've used a perpetual loss using high-level feature maps of the VGG network combined with a discriminator that encourages solutions perceptually hard to differentiate from the HR reference images. Different from previous works, we introduce a novel perpetual loss, colour loss to assist recover finer texture details.

This loss works to bound the mutations and learnings the network can do during the training period. It was observed during the training of the traditional SRGAN network that it has a tendency to diverge away from the original colours used to some different colours. Such a mutation can arise due to the fact that certain colour palettes can appear “real” to the discriminator and the VGG network might still be able to recognise similar features in such images. Hence a colour loss bound where the network can apply the gradients at the time of training. In our case, the colour loss is the binary cross-entropy error between the original image and the generated one. We also modify the existing deconvolution network to a convolution network and increase the number of residual blocks that showed a massive improvement.

MOTIVATION

A clearer image helps us experience objects, memories, explorations, in an unquestionably better manner. A high-resolution image is as close to what our eyes can see, as close to what we consider natural, so boundaries between what is man-made and what is natural starts to shake hands and vanish.

This brings our attention to the fact that a high-resolution image can be more than just a technical achievement. It can be the difference between an innocent being caught and let free, the difference between the discovery of a new star or identifying it as a fluke, and this is our primary motivator for this project where we aim to exploit the vast potential of deep neural networks, to create high-resolution images, from previously low-resolution ones.

Our motivation expands from the fact that this is a step towards not only helping people whose lives and work depend on high resolution images, but also towards solving the age old engineering problem of noise reduction and image enhancement.

TYPE OF PROJECT

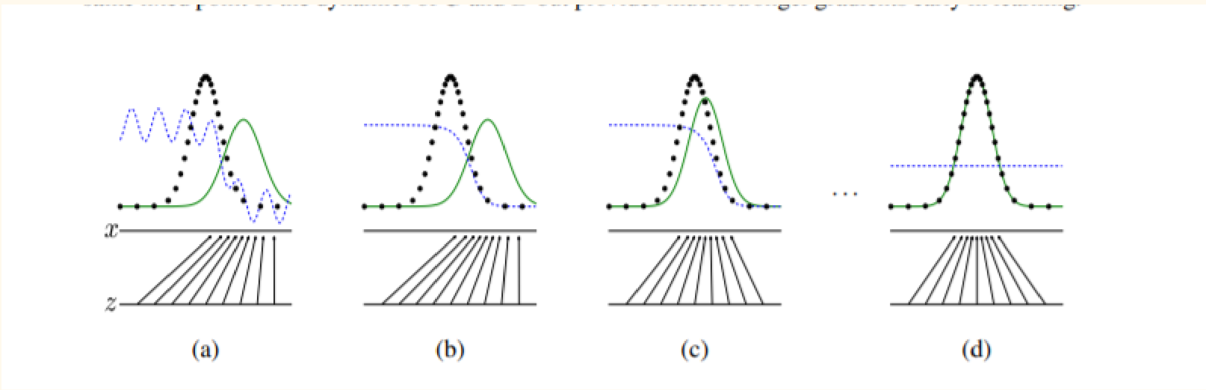
This project is a research cum development project. The area of super resolution is one of great research and such research usually takes a bit of time before seeing actual real world usage. However, new projects and startups are starting to deploy such ideas early on as experimental rollouts to the public. Hence this is also a development project on top of a research one.

LITERATURE REVIEW

We reviewed 6 different research papers to combine their methods, inspiration,ideas and outcomes to build up our project. The different literature reviewed is summarised sequentially.

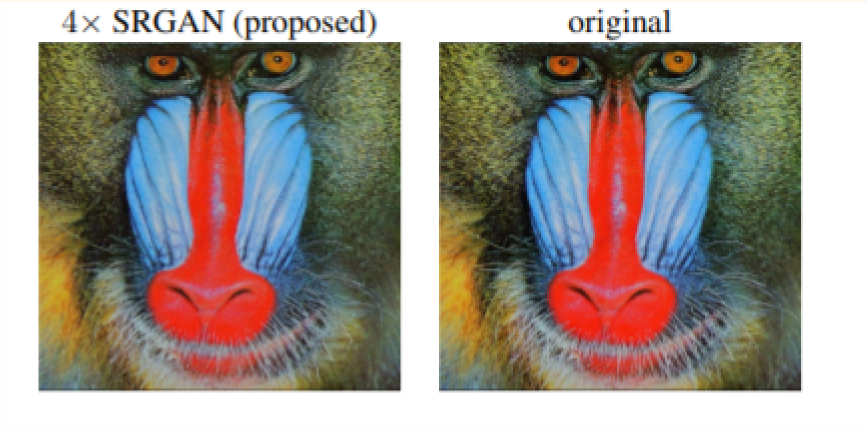
1.Generative Adversarial Nets by Ian J. Goodfellow [2]

In adversarial nets framework, the generative model is pitted against an adversary: a discriminative model that learns to determine whether a sample is from the model distribution or the data distribution. This framework can yield specific training algorithms for many kinds of model and optimization algorithm. The authors have explored the special case when the generative model generates samples by passing random noise through a multilayer perceptron, and the discriminative model is also a multilayer perceptron. We refer to this special case as adversarial nets.



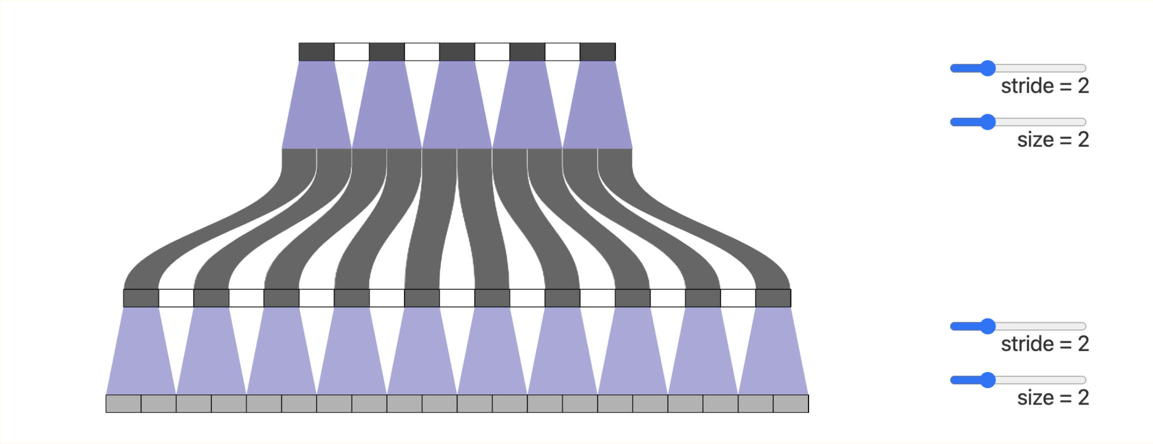
2.Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network -Christian Ledig [1]

The highly challenging task of estimating a highresolution (HR) image from its low-resolution (LR) counterpart is referred to as super-resolution (SR). SR received substantial attention from within the computer vision research community and has a wide range of applications



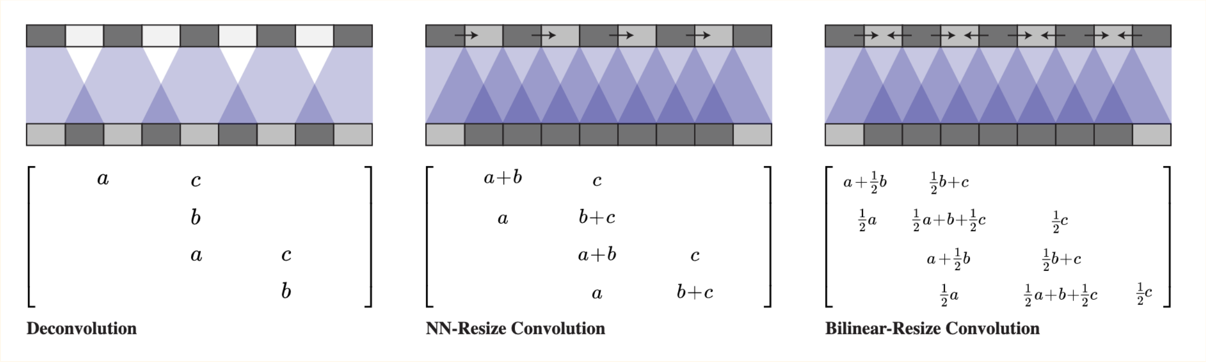
We took the concept of loss function to add to our discriminator network to enhance our work on the previous literature on generative adversarial network to minimise color loss and adversarial loss.

5.Deconvolution & Checkerboard Artifacts - Odena et al. - 2016 - [7]



Unfortunately, deconvolution can easily have “uneven overlap,” putting more of the metaphorical paint in some places than others. In particular, deconvolution has uneven overlap when the kernel size (the output window size) is not divisible by the stride (the spacing between points on the top).

Resize-convolution layers can be easily implemented in TensorFlow using tf.image.resize_images(). For best results, use tf.pad() before doing convolution with tf.nn.conv2d() to avoid boundary artifacts.



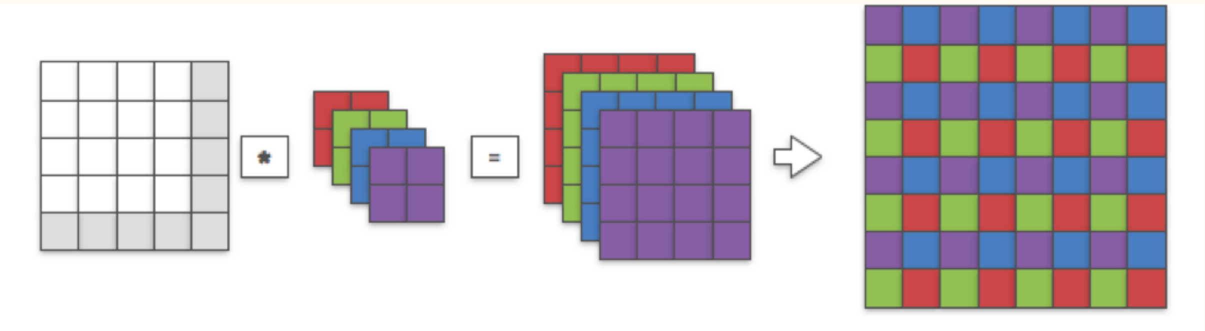
6.Checkerboard artifact free sub-pixel convolution - Andrew Aitken et al - [8]

Sub- pixel convolution is a specific implementation of a devolution layer that can be interpreted as a standard convolution in low- resolution space followed by a periodic shuffling operation.

Sub-pixel convolution has the advantage over standard resize convolutions that, at the same computational complexity, it has more parameters and thus better modelling power. Sub-pixel convolution is constrained to not allow deconvolution overlap , however it suffers from checkerboard artifacts following random initialization.

An initialization method for sub-pixel convolution known as convolution NN resize, compared to the sub-pixel convolution initialized with the schemes designed for standard convolution kernels, it is free from checkerboard artifacts immediately after initialization. Compared to resize convolution, at the same computational complexity, it has more modelling power and converges to solutions with smaller test errors.

The first two sources of artifacts deconvolution overlap and random initialization, can be eliminated by the resize convolution. Resize convolution first upscales the LR feature maps using nearest neighbour interpolation and then employs a standard convolutional layer with both input and output in HR space. Resize convolutions become a popular choice for generative modelling to alleviate checkerboard artifacts.



Method and Implementation

Our aim is to have a generating network represented by G that can minimise the low resolution artefacts and blurriness, or, to generate an HR image for an LR image that we input in. We aim to do this by training the network on a loss that is more akin to how humans perceive images and feature within thereof.

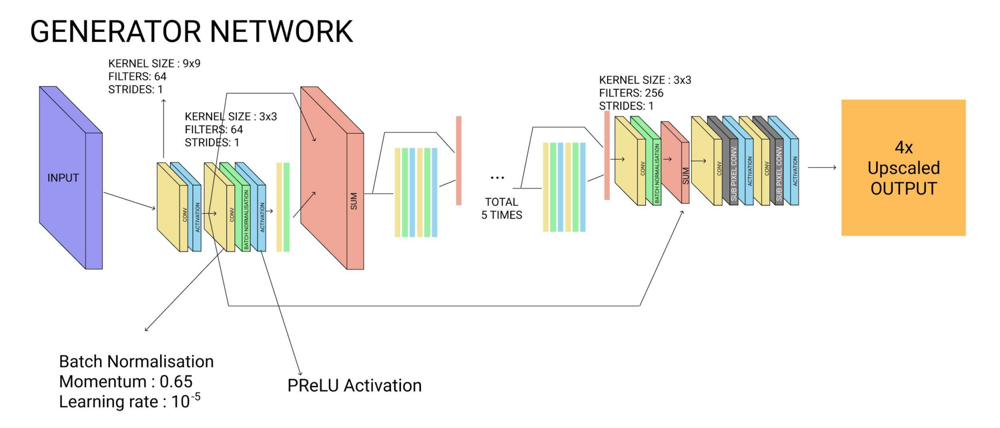
We aim to achieve this by using a feed forward deconvolution network as our generator Gsimilar to Ledig et al [1]. If we assume N total images to train on, $I_{HR,\,i}$ represent the ith HR image in the dataset, and $I_{LR,\,i}$ its LR counterpart. Assuming a loss function that closely represents our human perception of an image as L This means our goals remains to solve the equation:

$$\theta_{SR} \ = \ min(\frac{1}{N}\Sigma_{i=1}^{i=n}L \ (G(I_{LR,n}), \ I_{HR,n}))$$

This arbitrary loss function has been proposed by Ledig et al and called as perceptual loss. Here, after seeing problems in the said loss, we propose a new loss function. We name this colour bounded perceptual loss. We will discuss this further later.

Network Design

Building upon the general adversarial network building blocks of Goodfellow et al [2], we use the same ideology to work according to the classic minmax problem for adversarial networks.



This is how our generator network finally looks like. Pretty similar to Ledig et al with no substantial changes.

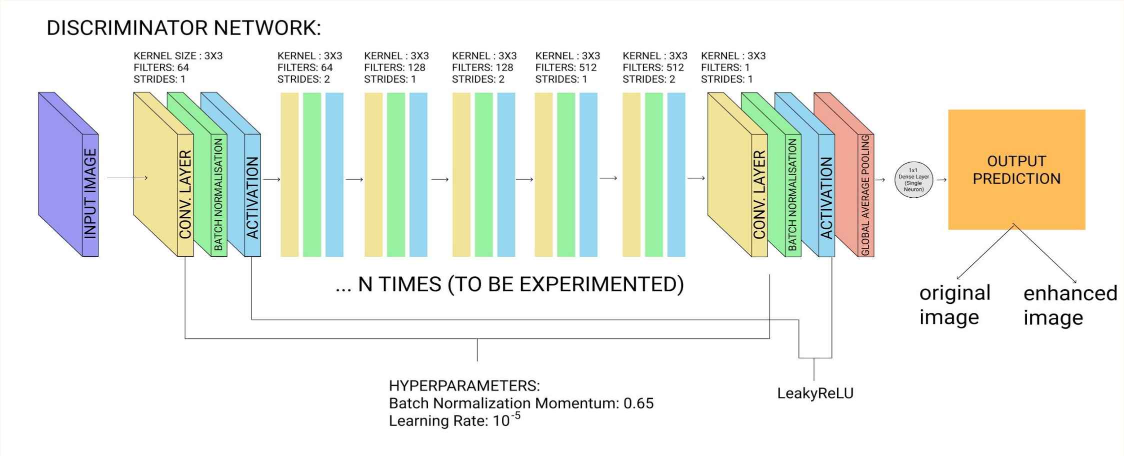
However, what has changed is the discriminator network. We took an approach which would let us use images of any size and resolution for the training of the discriminator network.

The Discriminator Architecture

To achieve this dimensional invariance, i.e, non dependence of the network on the size of input images, we use a pure convolutional network. The problem arises due to the fact that a dense layer requires a fixed number of inputs and outputs in order to function. Although the network described in Ledig et al[1] is convolution only, with the exception of a dense layer at the end. This dense layer forces you to have a fixed size image, which leads to an inferior learning.

Thus we propose a new discriminator model, working on top of the discriminator in Ledig et al[1], that does away with the dense networks. Instead what we use is a deconvolution layer with a single filter, so it produces a 2D matrix with all our predictions, followed by batch normalisation and activation. We pass this further into a Global Average Pooling layer, taking the idea from conah et al[4]. The global average pooling layer converts the matrix into a single value which we finally pass onto a single neuron with the idea that this distributes the weights of the average pooling layer for further biasing.

Along with this the network uses convolution, batch normalisation, and activation blocks with alternating 2x2 and 1x1 strides, and increasing kernel size as in the figure below.



Our discriminator network is now fully convolutional and works on inputs of any arbitrary size. We have hence achieved dimensional invariance with respect to training in both the generator and the discriminator of our network.

The next most important thing that remains to be defined is the loss function. Here, we use a modified version of the perceptual loss function as defined in Ledig et al[1] and call it a colour bounded perceptual loss function.

Colour Bounded Perceptual Loss

The main idea is to build a loss function different from traditional loss functions that produce a PSNR or SSIM score and update according to that. The idea is to create a loss function that can better represent how we as humans see and perceive photos. This was called the perceptual loss function in Ledig et al[1]. Ledig et al describes two kinds of losses : a VGG content loss and another, Adversarial loss. We observed gaps as these methods capture greatly how a human sees a photo, i.e, the VGG loss (discussed further below) does a great job of identifying features in a manner close to how we identify features such as lines and curves in our world, and the adversarial loss shows how the image deviates from reality. However, a gap was found as this does not include a loss for the colours used in the image. As was seen in training too, this lead to images that are not in the same colour shades as the original, however still seem like real images. Our final loss function looks like:

$$L_{SR} = L_{content} + 0.3\, L_{BCE}^{SR} + 10^{-3} L_{Gen}^{SR}$$

Colour Loss

The goal of this function is to create a loss value when the generated image deviates in the colours used by the CNN generator network. To achieve this, we devise this new colour loss function represented by L_{BCE}^{SR}. Here, we simply take a binary cross entropy on the generated and the HR images and use that as the loss, in combination with other losses. Thus, having N total images, and $I_G$as our generated image, our colour loss function would be:

$$L_{BCE}^{SR} =- \frac{1}{N} \left(\sum_{i=0}^{n} I_{HR,\, i} log(I_{G,\, i}) \right)$$

This represents with reasonable effectiveness, the deviations in colour of our generated images.

Content loss

Our content loss remains the same as in Ledig et al. It is the mean squared error between the outputs of the VGG 19 network by the group at oxford, for our generated image and the original HR image. This is given by:

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

Adversarial loss

Adversarial loss is the simple binary cross entropy loss between the prediction and reality. For generated images, it is the loss between the image and the value 0. For the original HR image, it is the loss between the image and the value 1.

$$l^{SR}_{Gen} = \sum_{n=1}^{N} - \log D_{\theta_D}(G_{\theta_G}(I^{LR}))$$

Residual Layers

SRGAN residual block is used in the SRGAN generator for image super resolution and uses a PReLU activation function to help training. We have changed the residual block layers from 5 to 16 in our model, and changed the activation from PReLU to LeakyReLU. This has shown to give better results.

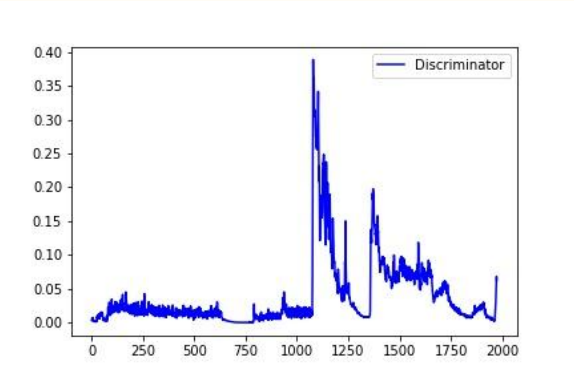
EXPERIMENTATION AND RESULTS

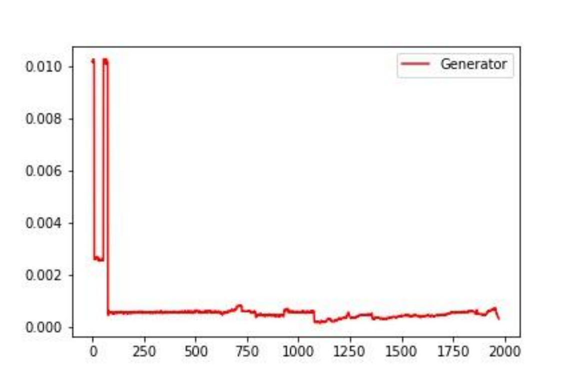
DATABASE

For the project, the DIV2K database was chosen. It contains 800 training images and 100 validation images. All images are of 2K HD resolution. These images were shuffled randomly, and sometimes cropped from random places, to create a variance in the dataset used. All of these 900 images have a low resolution counterpart in the dataset which is downscaled by a factor of 4 using bicubic interpolation. The dataset is the benchmark used in most Super resolution deep learning based approaches. It contains images of varying genres, including man made as well as natural objects, sceneries, water bodies, mountains, faces, houses, animals, food items, among others.

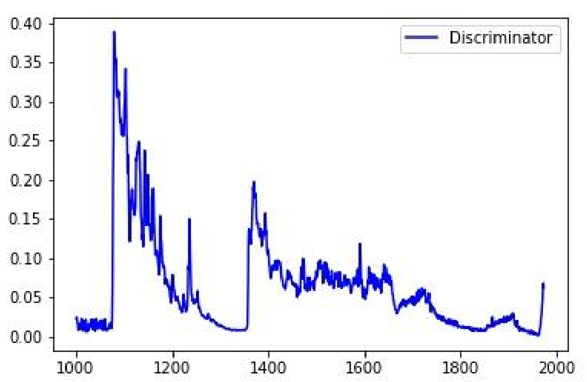
TRAINING DETAILS

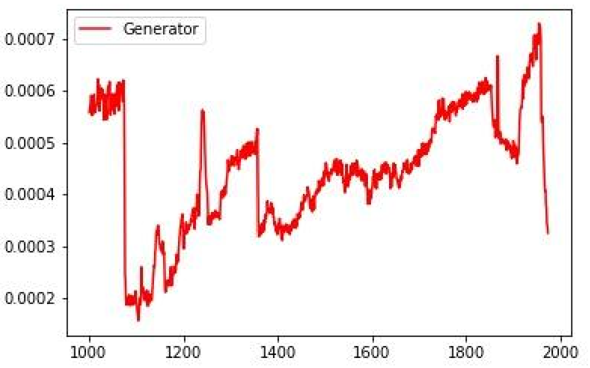
The network was trained on a google collaborator notebook. Mostly, this means it was trained on an NVIDIA T4 GPU with around 16GB VRAM and another 14 GB on computer RAM. The training was performed for a total of 2000 epochs. Two sets of models were trained in this. The first model, referred by network (1) used our colour bounded perceptual loss function for all the 2000 iterations. The other model, referred by network (2) used the original perceptual loss as defined by Ledig et al for the first ~1000 iterations and our colour bounded perceptual loss for the next 1000 iterations. It was settled on that the batch normalisation layers should use a momentum of 0.5 in both the discriminator and generator. The results however prove our network to be more on the unstable side and sometimes endures mutations that render it to a very high loss. However, after about 1000 iterations, a somewhat stability was observed. The learning rate chosen for the training was 10^{-5} in both the training models. The losses produced can be depicted here:





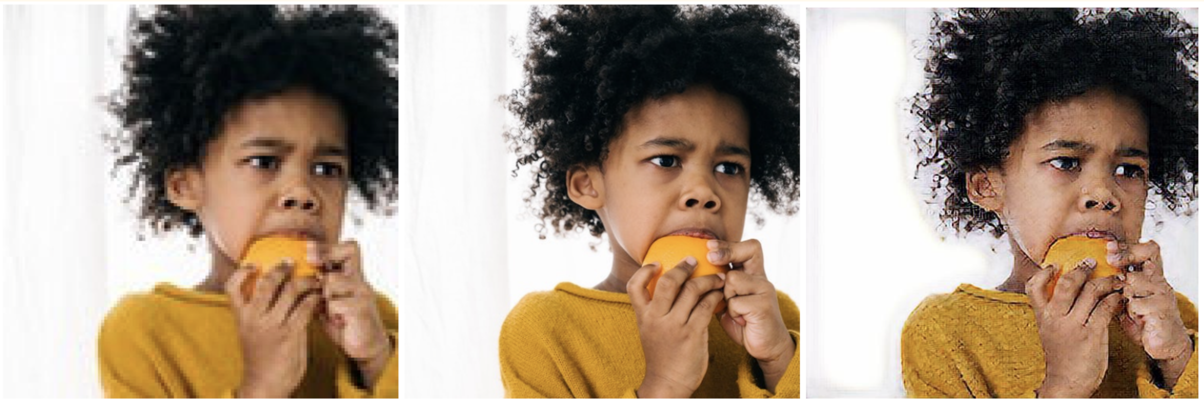
The losses after 1000 iterations, zoomed in, look like:





Opinion Testing

The LR images were upscaled using nearest neighbour and bicubic methods to a 4x upsampling. This, along with our generated images were shown to people to judge which looks better and clearer to a person, and in which, one can identify features and lines better. It was seen that almost all the times, the generated images were better in people's opinion. The images comparison looks like:



Low Scaled with Nearest Nbr

Original Image

Our Model

Comparison on traditional comparison benchmarks show these methods to not be the greatest improvement, but as is clear visually, the generated images are perceivably superior in almost all cases. Hence, the idea proposed in Ledig et al [1] is extended further that PSNR and SSIM are not the best benchmarks when it comes to realistic photo realistic super resolution.

Method	PSNR	SSIM
Nearest Neighbour	23.9647	0.6928
Proposed DCSRGAN	**24.6983**	**0.7073**

CONCLUSION

We have proposed improvements over the work of Ledig et al by 2 main new proposals. An introduction of a dimensional invariant discriminator network, consisting of pure convolutional layers, and a new colour loss function, to assist the general perceptual loss as in Ledig et al.

We also extend the idea tossed by Ledig et al that PSNR and SISM have limitations and are not the best when it comes to predicting comparative quality of an image. This idea was elaborated by an opinion testing, however, an extensive mean opinion testing could not be performed due to time and fund shortage.

Our project has achieved a 4x upscaling using two different training methods by a factor of 4x, and with popular opinion, creates more realistic and colour accurate images as compared to the original SRGAN. Not only this, but though the SRGAN was able to achieve this in about 2×10^5epochs, our network was able to achieve commendable results in only about 2000 epochs, which is a great improvement.

LANGUAGES USED

1. Python
   a. Tensorflow
   b. Numpy
   c. Matplotlib
   d. PIL
   e. Keras
   f. Tqdm
2. Google Colab notebooks
   a. Jupyter Notebook

NVIDIA GPUs + CUDA

CONTRIBUTIONS

- Rajat Goel 18103108 - Python Implementation with Tensorflow, finding colour loss, checkpointing
- Shreya Grover 18103113 - Understanding the project, devisign the methodology, finding colour loss and dimensional invariance
- Sarthak Sharma 18103112 - Literature review, suggesting Global Average Pooling, dimensional invariance

All members collaboratively worked in google colab notebooks to implement and improve the network and training. Inferences were also drawn collaboratively

REFERENCES

1. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network - Ledig et al
2. Generative Adversarial Nets - Ian J. Goodfellow et al
3. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network -Shit el al
4. Anysize GAN: A solution to the image-warping problem
5. An Introduction to Convolutional Neural Networks Keiron- O'Shea and Ryan Nas
6. Interpolation Techniques in Image Resampling - Manjunatha. S Malini M Patil
7. Odena, et al., "Deconvolution and Checkerboard Artifacts", Distill, 2016.
8. Checkerboard artifact free sub-pixel convolution - Andrew Aitken et al