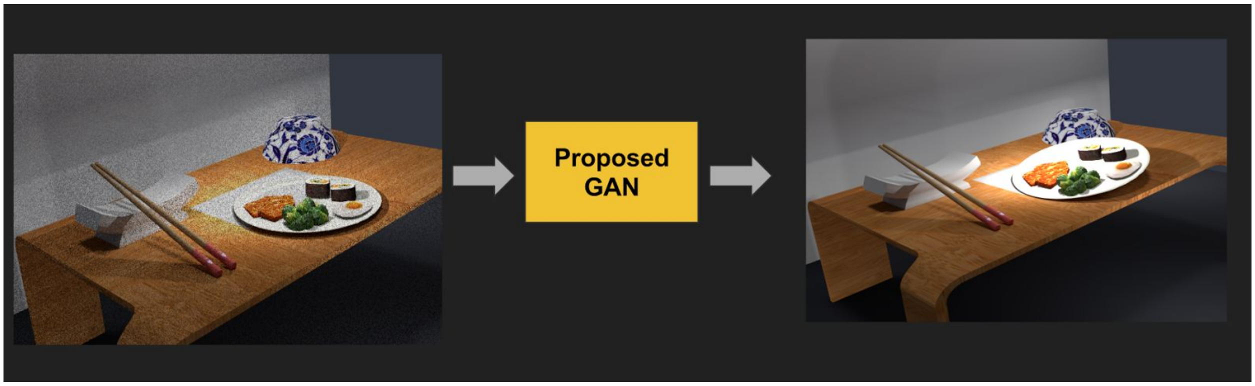
CS523 Multimedia Systems - Project 3

Project Proposal and Summary of the papers reviewed

Project Proposal:

Animation movie companies like Pixar and Dreamworks render their 3d scenes using a technique called Pathtracing which enables them to create high quality photorealistic frames. Pathtracing involves shooting 1000's of rays into a pixel randomly(Monte Carlo) which will then hit the objects in the scene and based on the reflective property of the object the rays reflect or refract or get absorbed. The colors returned by these rays are averaged to get the color of the pixel and this process is repeated for all the pixels. Due to the computational complexity it might take 8-16 hours to render a single frame.

We are proposing a neural network based solution for reducing 8-16 hours to a couple of seconds using a Generative Adversarial Network. The main idea behind this proposed method is to render using small number of samples per pixel (let say 4 spp or 8 spp instead of 32K ssp) and pass the noisy image to our network, which will generate a photorealistic image with high quality.



We are going to start with a base network(SRGAN or ID-CGAN) and use the loss function specified in semantic inpainting paper and then modify it based on our needs. The key for this project is the perceptual loss function. We have to define a refined perceptual loss which would preserve not only color and texture but also properties of the scene like motion blur and depth of field.

Papers reviewed by our group:

- Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network **- Abeer Alsaiari**
- Semantic Image Inpainting with Perceptual and Contextual Losses **- Ridhi Rustagi**
- Image De-raining using conditional generative adversarial network **- Manu Mathew Thomas**

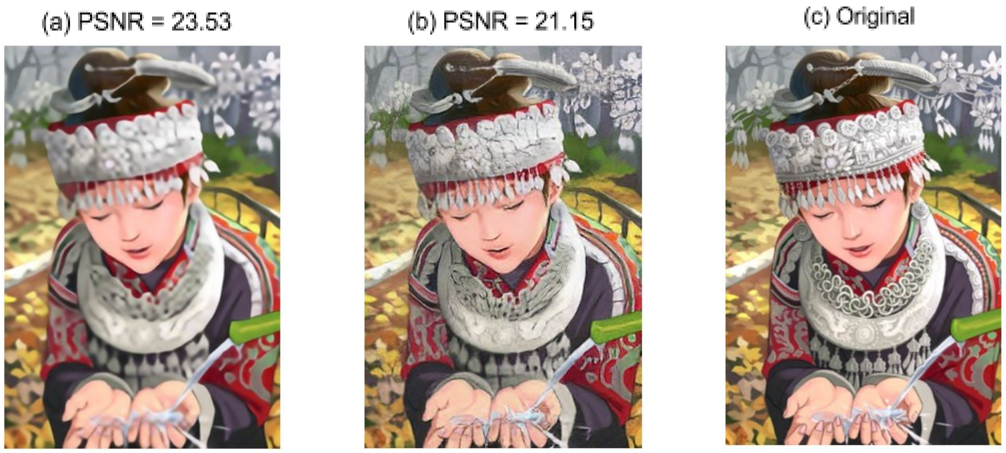
Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network

https://arxiv.org/pdf/1609.04802.pdf

In this paper (Ledig et al., 2016), the authors proposed a Super-Resolution Generative Adversarial Network that resolves perceptually satisfying high-resolution image from its low-resolution counterpart. They proposed a new loss function that handles the problem of lost fine texture details in the reconstructed image. Using a loss function based on high-level feature maps of the VGG network, the reconstructed image is more realistic and perceptually satisfying.

Problem Statement

Most previous super-resolution algorithms are based on Mean Squared Error (MSE) as an optimization measure. The super-resolution algorithm converges to the optimal solution by minimizing the MSE between the resolved image and the reference image. This optimization measure is defined on the pixel space. The model pushes to find a solution by averaging possible solutions in the pixel space. Hence, it produces an overly smooth output image, which lacks fine details such as texture. Therefore, MSE-based super-resolution algorithms cannot handle the recovering of high-frequency details and provides poor perceptual quality. Minimizing MSE also means maximizing peak signal-to-noise ratio (PSNR). This ratio is typically used to measure the quality of the super-resolution algorithm. It is a quality measurement between the resolved and the original image. Higher value of PSNR is desirable because it means that the ratio of original image to reconstruction error is higher. The goal of previous super-resolution algorithms is to minimize MSE and hence to maximize PSNR. However, highest value of PSNR doesn't necessarily yield a perceptually satisfying image. For example, image (b) in Fig. 1 is more realistic and perceptually satisfying although it has less PSNR value than image (a). The goal of the proposed Super-Resolution Generative Adversarial Network (SRGAN) is to recover the photo-realistic textures in the reconstructed image.



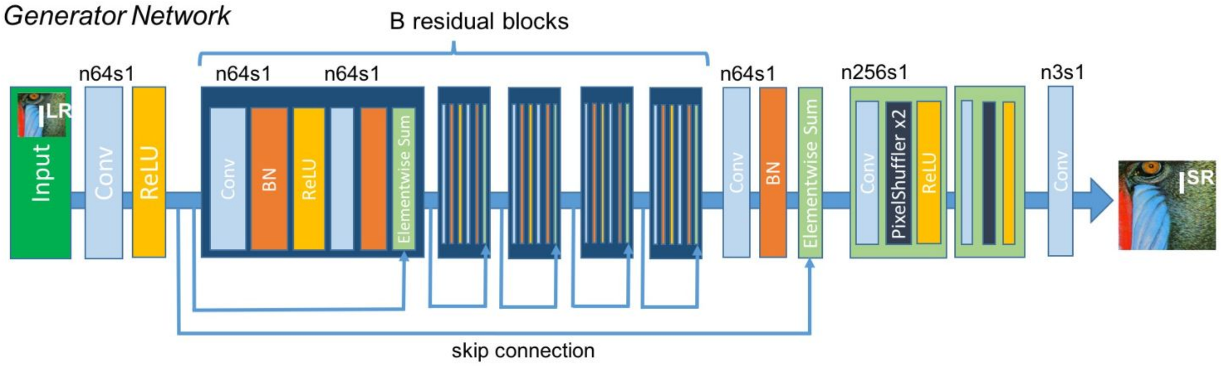
(a) PSNR = 23.53

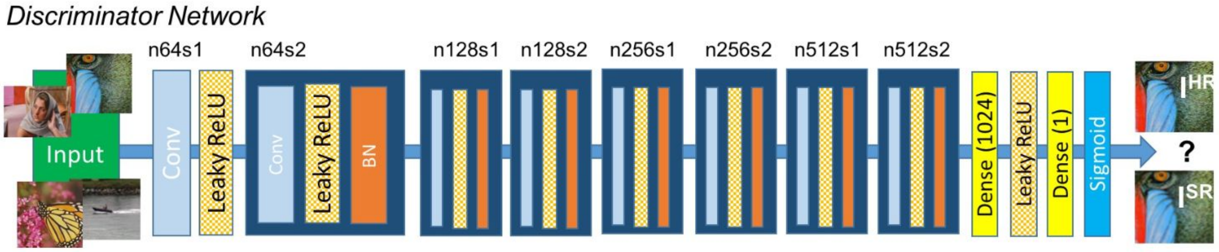
(b) PSNR = 21.15

(c) Original

Method

In this paper, they presented a Super-Resolution Generative Adversarial Network (SRGAN) that is capable of inferring more realistic, natural and perceptually satisfying images. They proposed a very deep residual net architecture, which is a GAN-based network for single image super resolution. The architecture consists of Discriminator and Generator networks. Both networks are trained by minimizing a perceptual loss function as defined below. The network takes for training a set of low-resolution images and their counterpart high-resolution images; and trains the generator to generate an indistinguishable image from the ground truth. So, it fools the discriminator with the reconstructed image. Similarly, the discriminator is trained to discriminate the reconstructed image from the real image. The generator network consists of a number of residual blocks each of which consists of 2 convolutional layers followed by batch-normalization layers. The discriminator consists of eight convolutional layers followed by two dense layers and a sigmoid activation to produce the probability of the output classification. The training of the network is achieved by minimizing the perceptual loss function, that is, minimizing the weighted sum of its components: content loss and adversarial loss.





Perceptual loss function

Instead of relying on pixel-wise error measures such as MSE-based optimization, they proposed a novel perceptual loss function consisting of content loss and adversarial loss. The goal of adding content loss is to handle the solution with respect to perceptual high-level features. The content loss is based on perceptual similarity. Using pre-trained 19-layers VGG network, feature maps are obtained for the reconstructed and reference images. The feature map is computed by encoding each image vector by layer filters. The difference between features maps of reconstructed images and reference images is computed as a Euclidean distance to define the content loss. The adversarial loss makes the discriminator and generator push the solution to the natural image space in which the generator tries to fool the discriminator with the reconstructed image and the discriminator tries to distinguish the reconstructed image from the real image.

Dataset

They did experiments on three widely used dataset from three different papers; each of which is a collection of high-resolution images and their corresponding blurred and downscaled as low-resolution samples.

Semantic Image Inpainting with Perceptual and Contextual Losses

arxiv.org/pdf/1607.07539.pdf

Main idea

The goal of inpainting is to reconstruct the missing or damaged portions of an image. It has numerous applications such as restoration of damaged paintings or image editing. In this paper the authors propose a novel method for image inpainting based on a Deep Convolutional Generative Adversarial Network (DCGAN). Given a corrupted image, back propagation is done on contextual and perceptual loss to map the corrupted image to a smaller latent space. The mapped vector is then passed through the generative model to predict the missing content.

Previous Work

Some of the previous work done in the domain of image inpainting rely on prior information that exist in input image such as hole filling using the nearest patch information and reconstruction of image given the planar and/or low rank structure of the input image. These local methods can be efficient and achieve high quality results when the above assumptions are satisfied. While they are good at seamlessly filling holes with local pixels/patches, they are unable to predict the semantic information in the missing region, especially when the desired content is not contained in the corrupted image.

To solve the more general inpainting problem, non-local methods which use the cut/paste technique were used but they also failed when the scene is not contained in the training set. To overcome this failure, sparse coding was proposed to recover the images from a learned dictionary and recently, inpainting by convolutional neural networks has shown promising results. However, the authors are interested in the more general and more difficult task called semantic inpainting which aims to predict the detailed contents in a large region based on the context of surrounding pixels.

Innovation

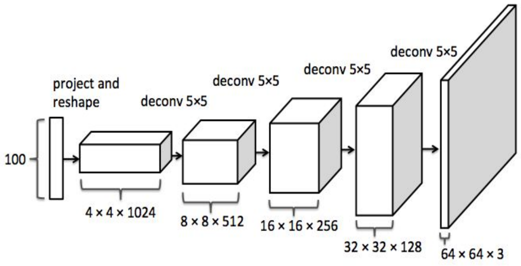
This paper considers semantic inpainting as a constrained image generation problem. The authors propose a new framework to fill holes in images using back-propagation on a pretrained image generative model, such as the Deep Convolutional Generative Adversarial Network (DCGAN). With the corrupted image, the proposed method hopes to find the "closest" corresponding vector in the smaller latent space, which is then used to reconstruct the whole image. To find the "closest" vector in the latent space backpropagation is applied on the designed loss function. The paper defines two losses : 1. contextual loss to limit deviation between the recovered image and the corrupted image 2. perceptual loss which penalizes recovered images that are perceptually unrealistic. The proposed framework can be used to fill arbitrary holes in the image without retraining the network.

Method

The DCGAN model architecture from Radford et al is used in this work .

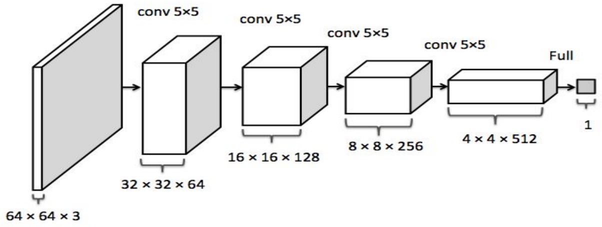
Architecture

Generative Model G is designed as follows:



The **input** is a 100 dimensional random noise vector, drawn from uniform distribution
between [1; 1] , followed by an 8192 fully connected layer reshaped into dimensions of 44512 .
Each of the following layers are deconvolutional layers, where the numbers of channels are halved, and image dimension doubles from the previous layers. The **output** layer is of dimension 64643 , the size of the image space.

Discriminative model D is designed as follows:

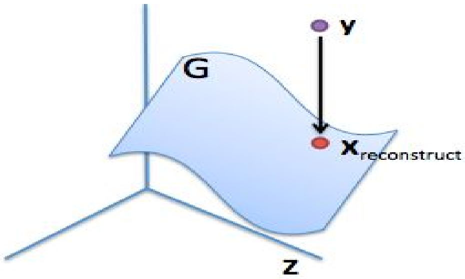


The input layer is an image of dimension 64 X 64 X 3 , followed by a series of convolution layers where the image dimension is halved, and the number of channels is doubled from the previous layer, and the output layer is a two class softmax.

Inpainting task:

- G maps a random vector z, sampled from a prior distribution pZ, to the image space. D maps an input image to a likelihood. The purpose of G is to generate realistic images, while D plays an adversarial role to discriminate between the image generated from G, and the image sampled from data distribution pdata.
- The networks are trained by optimizing the loss function.

- After training, G is able to embed the images from pdata onto some non-linear manifold of z .



- Corrupted image is denoted as *y*
- To do reconstruction, we hope to recover the “closest” image on the manifold to the corrupted image.

- **Contextual Loss** : It is used to measure the context similarity between the reconstructed image and the uncorrupted portion.

  $$\mathcal{L}_{contextual}(\mathbf{z}) = \|\mathbf{M} \odot G(\mathbf{z}) - \mathbf{M} \odot \mathbf{y}\|_1$$

  - M is the binary mask of uncorruption.
- **Perceptual Loss** : It encourages the reconstructed image to be similar to the samples drawn from the training set.

  $$\mathcal{L}_{perceptual}(\mathbf{z}) = \log(1 - D(G(\mathbf{z})))$$

With the above defined losses the corrupted image can be mapped to the closest z in the latent representation space. z is updated using back-propagation with the **total loss**:

$$\hat{\mathbf{z}} = \arg\min_{\mathbf{z}}(\mathcal{L}_{contextual}(\mathbf{z}) + \lambda\mathcal{L}_{perceptual}(\mathbf{z}))$$

- Poisson blending is used to reconstruct the final results.

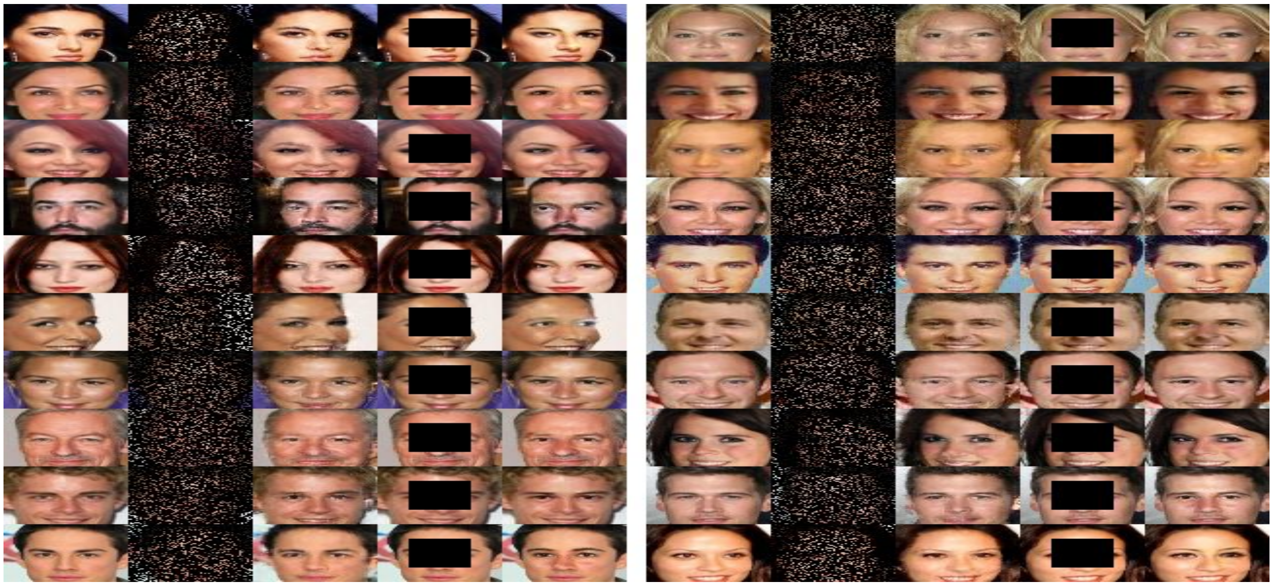
Dataset and Results

1. CelebA dataset

The results are evalutaed on the CelebA dataset, with two types of corruptions: 80% of pixels randomly deleted, or a large missing block in the central.

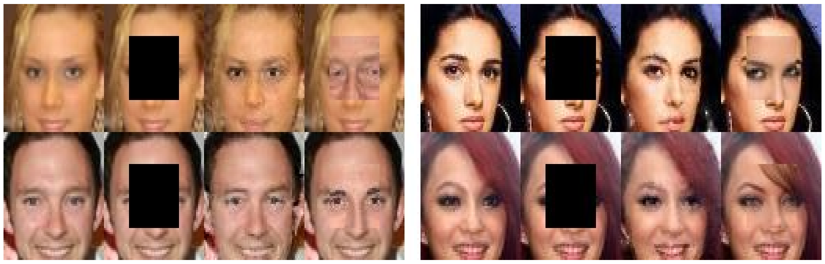
2. SVHN Dataset Street View House Numbers

The (SVHN) dataset contains a total of 99,289 RGB images of cropped house numbers. Input images are resize to 64 X 64 to fit the DCGAN model architecture.



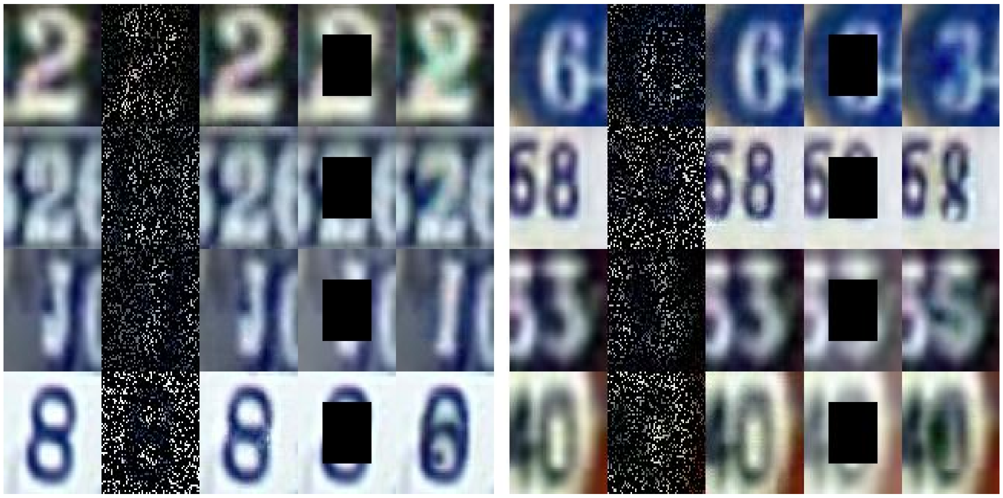
Column 1: Original images from the dataset. Column 2: Images with 80% random missing pixels. Column 3: Inpainting of column 2 by our method. Column 4: Image with large central region missing. Column 5: Inpainting of column 4 by this paper’s method.

Comparing with another method:



Central hole filling. For each example, Column 1: Original images from the dataset. Column 2: The corrupted image. Column 3: Inpainting by our method. Column 4: Inpainting by the nearest patch in the dataset.

SVHN dataset results:



Column 1: Original images from the dataset. Column 2: Images with 80% random missing pixels. Column 3: Inpainting of column 2 by our method.

Column4: Image with large central region missing. Column 5: Inpainting of column 4 by our method

Conclusion

The above results clearly show that the proposed method using DCGAN produces semantically photo realistic images than any other CNN based method.

Image De-raining using a conditional generative adversarial network

https://arxiv.org/pdf/1701.05957v3.pdf

The main idea behind this paper is to increase the visual quality of a rainy image by removing rain streaks from the image using neural networks. The authors of the paper He Zhang et al. demonstrates a simple technique using a conditional generative adversarial network to remove rain streaks from images which surpasses all other state of the art techniques.

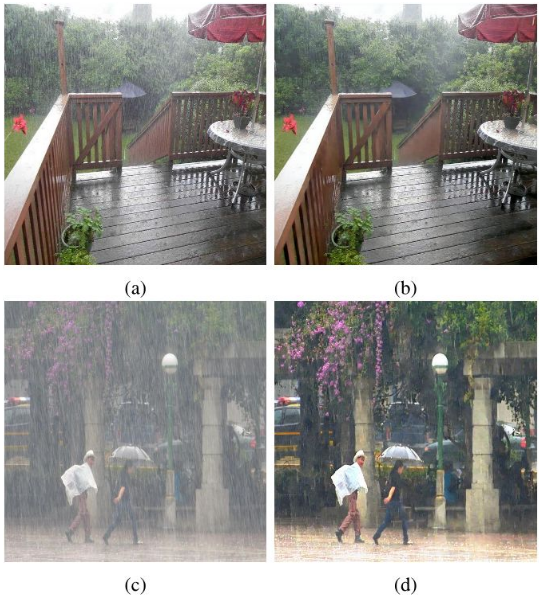
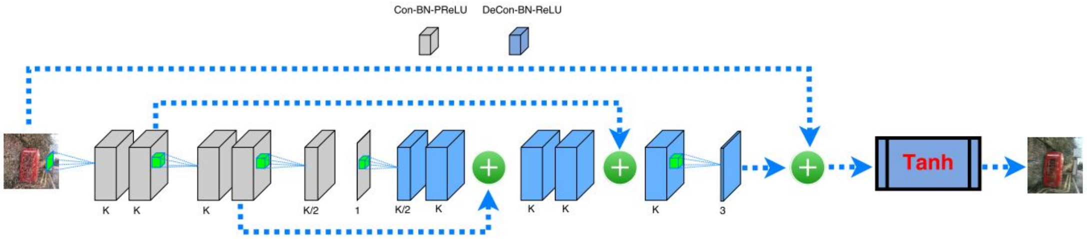


Fig. 1: Sample results of the proposed ID-CGAN method for single image de-raining. (a)&(c) Input rainy images. (b)&(d) De-rained results.

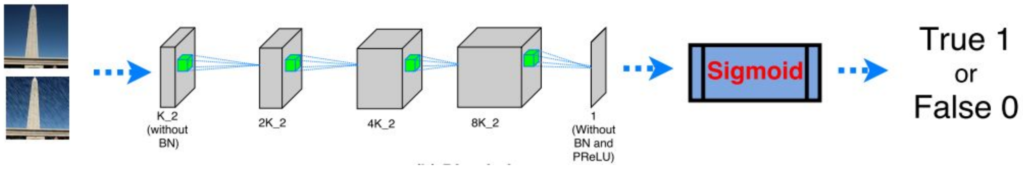
The network consist of a 1) Generator, whose job is to take a rainy image as the input and generate an output image with all the rain streaks removed 2) Discriminator, who is responsible for determining whether the input given to it is a real image or a synthesised image 3) Refined perceptual loss function, which not only maintains the high level feature like all other perceptual loss but also preserves the color and texture informations.

Generator Network

The generator network consist of 5 convolutional layers with number of filters K and K/2 (authors used K = 64). Each convolutional layer consist of convolution, batch normalization and PReLU activation function (which is same leaky ReLU). Convolution layers are followed by deconvolution layers which is basically a mirror image of the convolutional layers. Skip connections are used every 2 conv layers, it's adding up to its deconv counterpart to recover the spatial information. Finally an activation function Tanh is used to get the output image. One interesting to note here is that the generator doesn't use any kind of pooling like max pooling or fully connected layers.



Discriminator Network



The discriminator network consist of 4 convolution layers with number of filter K, 2K, 4K and 8K(author used K = 48). Discriminator takes an input and determines whether its a real image or a fake image. This make sure that the generator with help of the refined perceptual loss function to generate more convincing images to fool the discriminator.

Refined Perceptual Loss Function

Perceptual loss function generally maintains the high level features of an image, but here color and texture is also important and needs to be preserved. Authors of this paper proposed a new loss function based on the general perceptual loss function and is defined as:

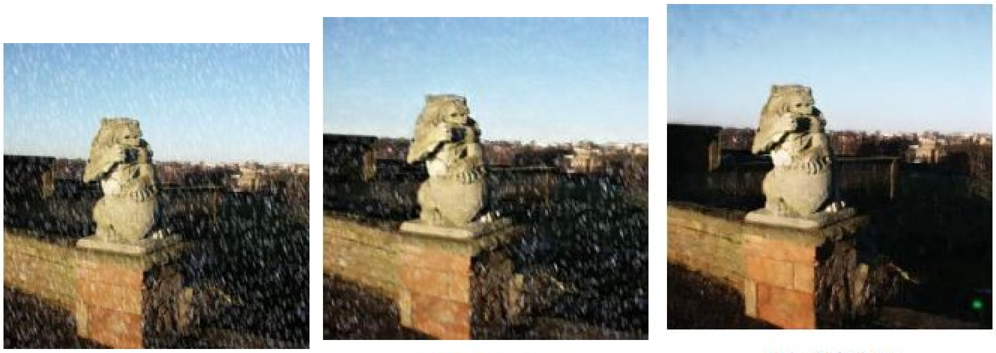
$$\mathbf{L_{RP} = L_E + \lambda_a L_A + \lambda_p L_P}$$

Where $\mathbf{L_E}$ is per pixel loss function called Euclidean loss, $\mathbf{L_A}$ adversarial loss from the discriminator and $\mathbf{L_P}$ perceptual loss calculated from relu2_2 in VGG-16 model. $\mathbf{\Lambda_a}$ and $\mathbf{\lambda_p}$ are pre-defined weights for adversarial loss and perceptual loss.

Dataset

Due to the lack of availability of large dataset with rainy images, authors extracted 700 images from UCID dataset and BSD-500 dataset. Then they added rain streaks using photoshop. Only 50 real rainy images were used and is only used for evaluation.

Results



Input CNN [12] ID-CGAN

First image is the rainy input, second is a CNN based rain streak removal technique and third is the technique explained in the paper. Cleary, ID-CGAN performs much better than CNN based technique.