Generative Adversarial Networks for Facial Image Inpainting and Super-Resolution

Aishwarya G\textsuperscript{1,2} and Raghesh Krishnan K\textsuperscript{1,3}

\textsuperscript{1}Dept of Computer Science and Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India  
\textsuperscript{2}cb.en.p2aid19002@cb.students.amrita.edu  
\textsuperscript{3}k_raghesh@cb.amrita.edu

Abstract. Inpainting helps to fill in the lost data in visual images. Inpainting techniques also refer to unusual image editing in distorted regions. These include areas that are noisy, blurred and watery areas. The most appropriate pixel values must be replaced in these regions to achieve good performance. Artists used to play it, and still now, pieces that are not in the picture can be inpainted in the same manner, though it takes more time. In the present age of automation, inpainting can be automated to obtain quicker and better outcomes by deep learning technologies. In this area, many of the latest techniques have been created, however, many methods produce blurred findings and data loss. Two adversarial networks are used to achieve this task, where first network aims at inpainting and the second network aims at super-resolution. The input generated as a part of first stage network is passed on to the second stage super-resolution network to overcome blurriness that is caused in the initial inpainting network. The network efficiency is determined in terms of increased PSNR obtained which is 28.19 dB with less training period of approximately 14 hours in comparison with other network models which performs similar task.

1. Introduction

Image inpainting is a form of preservation and image restoration method, initially proposed by Pietro Edwards, director of the Restoration of the Public Pictures in Venice, Italy to preserve famous works[1]. Recent technological advances have greatly improved inpainting, which helps to restore and repair old and damaged images. Previously, image painting was done using diffusion-based methods that spread the local element in unknown regions or model-based methods that reconstructed a lost part such as a single pixel or a pixel at a time, while keeping in contact with neighboring pixels. The above methods fail when size is too large or complex; hence there is a need for something more that provides practical thinking (visualization). With the increasing demand for high quality images, the need to rebuild damaged image areas is growing exponentially. Furthermore, since there is no learning involved with these conventional approaches, they do not understand the meaning of the field. Accessible approaches rely on nearby pixels to fill in incomplete circuits and may not be applicable in all cases. Comprehensive learning methods resolve this issue by locating the appropriate pixels to fill
in non-contextual regions. Face inpainting, also renowned as facial completion, provides a clear and realistic facial inpainting for an image that has missing or distorted region. Faces play an inevitable role in representing human characters, the foundation of facial authentication and identification is formed because of facial painting when the destruction is seen on the face part of the image. The biggest challenge in face painting is that the model must fill in the missing space with the appropriate facial features, otherwise the result can be distorted. Therefore, this study is focused on face inpainting with super-resolution to get high quality visual images as it is very important in today's computer vision.

2. Related works

Image inpainting is considered to be one of the most exciting areas of research in recent trends. The approach of Wasserstein GAN proposed by Yi Jang et al., [2] uses skip connection which is embedded to the generative network after encoder architecture which will increase the generated models prediction potential. The data sets used for this research are LFW and Celebrity Face Image Dataset. The results obtained in these research works are blurred images which is caused due to structure-based synthesis [3]. The image inpainting model by Ugur et al., [4] uses a combination of lower level merging and higher level splitting of a technique based on patch such as PatchGan and general GAN can contribute in accomplishing local continuity after image while it confirms the holistic nature of the image. This research has been demonstrated on street data sets and location based data sets such as Paris street view and Imagenet along with celebrity face image data set. Yi Wang et al.[5] proposed architecture was optimized to replicate the original image in terms of quality and visualization using ID-MRF and a multilayer neural network which had different structural networks than the normally used networks.

Raymond A Leht et al.’s general DCGAN[6] architecture implements Deep Convolutional Generative network where the number of convolutional layers are increased in number. The generative network inputs a random standard dimensional vector with a uniform distribution between $[-1, 1]$ and produces a standard sized image. The discriminator model, $D$, is employed in reverse order of the generative model. Just after input layer, there is a series of convolutional layers that minimize the image dimension twice. The channel numbers are twice the size of the preceding layer and two-class softmax is used for output layer. The model outperformed Context Encoders for inpainting the missing regions. Zhao Gao et. al[17], in their research work proposed a method using CNN for increasing the resolution of the image implementing self-feature loss. Training took three days to obtain good results in the DIV2K dataset.

The main aim of using feature extraction implementing auto encoder decoder-based model[15] is to acquire the right set of features through learning to predict in uncertain situations such as an image with larger missing region. A two-stage generator was used initially to reconstruct the image and the output is passed to obtain super resolution images in fine grained network proposed by YangLi et al [7]. Hang Wang et al, proposed an enhanced encoder network which uses parametric RELU for training on landscape datasets that obtains different set of features[12]. Wiemin Tan et al[8], proposed a three-block model, that has a additional feature extraction layer which aimed at providing good visual results. This concentrated on various orientation attributes that constitutes to learning with imbalance in data[9].

The GAN network architecture consists of convolutional layers with sub-sampling layer where upsampling and downsampling of the image pixels is done[10]. The researchers concentrated on many datasets that ideally included face images and also location based datasets. Most of the works implemented Wasserstein GAN which maintains a constant phase between generating and
discriminating of the images. The above mentioned works aimed at providing good quality inpainting results that serves the purpose of appropriate regeneration of pixels. The face image inpainting task consumes a lot of training period and the models failed when inpainting side profile orientation of the face images as it involved many convolution and pooling layers. The proposed architecture in this paper, overcome this drawback from previously proposed architecture by using appropriate layers for convolution and use L1 loss to contribute faster training with no compromise in outputs obtained in comparison to PSNR values obtained using previously proposed methods.

3. Methodology
Generative Adversarial Networks is a challenge to train. In the architecture used to inpaint human face images shown in Figure 1, the discriminator attempts to detect how far the ground truth image varies from the generated image. The image to be inpainted is sent to a generator, over time the real image is compared to the bias and the variance is determined according to the loss functions, and the model is later updated with weights to get the right collection of parameter values to execute this function. The discriminator model uses the sigmoid function to operate in the output extraction layer to detect the binary split problem and suits the function of the binary cross-entropy loss function. The value of the input given is correct (or incorrect as one under the prediction) as 0 to 1 is in turn predicted by this model. The loss function penalizes the model for how it varies from the anticipated spread of a single initial image in the anticipated distribution. This is a description of the error reported by the discriminator and generator in the following batch of training.

![Image Inpainting using GAN.](image)

In addition to the above network, in the architecture of super-resolution network, VGG19 model is used in generator phase to extract features of the image followed by same layers with inpainted image as input and high resolution image as output in the super resolution.

3.1 Dataset and Preprocessing
The CelebA dataset provided by MMLAB[11] for non-commercial purposes is used in this work. The dataset consists of 2,00,000 celebrity images which differs in terms of various attributes such as hair, nose, glasses, side profile etc,. The dimension of these images is 125x125. The same dataset is used in both the inpainting and super-resolution network. With the change in the training perspective, where two networks has to perform different tasks, hence the preprocessing is carried out differently. For the inpainting network, the image is first converted into a tensor and the image is masked with size 64x64
for centre mask and random mask can also be designed to cover random regions. In this case, the input image will be the masked image, the expected output will be the ground truth image and the generator network will provide the output image based on the balanced training with the discriminator. Similarly, for the super-resolution network, the images are first converted to tensor and the bipolar interpolation is applied to the image for training to pass as input to the network. In the super-resolution network, the input will be bipolar interpolated image, the expected output is original image and the high resolution image will be provided as a output from the generator. The models were trained on NVIDIA P100 GPU and 16 GB GPU Memory. It can be observed that GAN networks generates new data that resembles closer with the original data based on the loss values of discriminator during training.

3.2 Inpainting Using GAN

Inpainting methods are of two types. The conventional one does not know any information about the pixels in the missing area, which implies that the model is blind. The other is a non-blind method which knows that the information has to be in painted in the particular area based upon the knowledge gained on learning [12]. The challenge in inpainting faced by many researches is when the size of the mask grows, the network struggles to fill the region accurately. The region masks can be of any type, centre mask or random masks. These masks can be given the same pixel values, such as 1 or 0 or it can also be some randomly generated noise. Though GANs produced promising results for generating missing regions, it is equally challenging to train them. GANs are used in tasks such as image restoration, generation and translation. The region masks must be properly filled to produce the highest appropriate result using proper training.

As discussed in Section 3, the GAN network consists of, a generator and a discriminator. The generation phase produces a different set of data values and tries to deceive discrimination. It contains CNN layers that are followed by an activation function like Sigmoid which provides the required output at a single value. The loss function is divided into two parts namely, the reconstruction loss which is the L2 loss function and the generation loss. The formation and cohesion of a lost area according to its meaning can be captured through this loss as shown in (1),

$$\text{L}_{\text{reconstruct}} = || \hat{M} \odot (x – F (1 - \hat{M}) \odot x) ||^2_2$$  \hspace{1cm} (1)

where, $\text{L}_{\text{reconstruct}}$ represents the reconstruction loss, $\hat{M}$ is the retrieved image, $F$ is a two-dimensional image function and $x$ is the pixel element of the image.

It is necessary to understand that use of L2 alone would give us a distorted image here because it reduces the mean value. The second loss function used is adversarial loss which attempts to make the predicted output 'resemble' real, and this helps to overcome the blurry image that the L2 loss might have produced. It can be expressed as given in equation (2),

$$\text{L}_{\text{adversarial}} = \max \mathbb{D} \mathbb{E} x \in X [\log (D(x))] + \mathbb{E} z \in Z [\log (1 - (D(G(z))))]$$  \hspace{1cm} (2)

where, adversarial loss $\text{L}_{\text{adversarial}}$, inclines to make the generated image to look like the real one and not just the certain region and the network tends to give realistic look to the image. D is the discriminator and G is the generator function. The total loss is arrived as per equation (3),

$$\text{Total loss } L = \lambda_{\text{reconstruct}} \cdot \text{L}_{\text{reconstruct}} + \lambda_{\text{adversarial}} \cdot \text{L}_{\text{adversarial}}$$  \hspace{1cm} (3)
The $L_{\text{adversarial}}$ and $L_{\text{reconstruct}}$ loss function in (3) is tuned with hyper parameter $\lambda$, inorder to give importance to the adversarial and reconstruction losses.

### 3.3 Super-resolution Using GAN

The super-resolution network aims to provide higher resolution of the image that is produced as an output of initial inpainting network. The super-resolution network takes input as original image and uses bicubic interpolation to get lower resolution image. The high resolution images are formed by using transforms function and converted to torch tensors similar to that of creating lower resolution images but does not change the original pixel values. A VGG19 network is used for feature extraction which can be used by the generator network[16] in order to learn the features of the input image more precisely. Based on the loss function value that is determined the between the generated features and the original image features the network provides super resolution images.

To reduce computation and blurriness, L1 loss is used for reconstruction and MSE for pixel-wise discrimination as given in equation (4) and Mean Squared Error is calculated by taking the average value square of L1 loss respectively for both inpainting and super resolution of images.

$$L_1 \text{ Loss Function } = \Sigma_{i=1}^{n} |y_{\text{true}} - y_{\text{predicted}}|$$  

(4)

In-depth learning models can take more time train them; it is important to save and upload models for the later use and faster recovery of parameters and weights. The weights obtained during training of the network are saved in the form of h5py files. Weights are stored using and later uploaded using the torch function. The stored model loads and network parameters are loaded and the model is rebuilt, and the model is assembled before use. The model is tested directly to print the same kind of face images data to get accurate and high resolution images, and then tested with standard metrics. The entire overview of the system is represented in Figure 2.

![Figure 2. Design Overview](image)

### 4. Results

GANs are a type of unsupervised learning algorithms. Therefore, the results obtained cannot be tested guaranteed with accuracy as it cannot be compared with actual value[13]. With little loss, it can be determined that the model has learned to inpaint the regions well. PSNR values may sometime indicate
poor results as GANs are used for generating new data, but in order to compare with the existing methods PSNR values are obtained.

The pixel wise loss to every image is determined using the mean squared error. After some time, the amount of loss decreases. The model is operated for 50 epochs, approximately 9 hours to execute. A graphical plot indicating the mean square error of each image on average is given in Figure 3.

![Figure 3. MSE vs Images](image)

The median results obtained for the PSNR value are approximately 65.17 dB. Values are printed whenever a collection is divided by 500 for visual purposes. Weight recovery takes place to obtain the appropriate parameters for the inpainting application and the model performs the task.

The result of this phase is as follows in Figure 4. The first row constitutes a masked image with size 64x64, the next represents the output image, and the last represents the ground truth. This ground truth image is passed on to discriminator to identify the real image or image created, depending on when the loss of discriminator is obtained.

![Figure 4. Output Visualizations](image)

At the beginning of the training, the discriminator loss is even greater. This shows that the painted image does not match the original image. Generator losses should be greater while Discriminator losses are reduced which can be noted in subsequent periods. Discriminator losses are expected to be lower than Generator loss inorder to get high quality inpaint results. The same training was performed for the super resolution network and subsequent loss of value was observed during training is given in Figure 5. It can be observed that the loss values are very minimal for each image which provides an insight that the training of the model is efficient to obtain good results. Though the results obtained are
good in terms of structural orientation, blurriness is visible in generated images due to use of L1. And if L2 loss is used we will witness structural disorientations in image as it does not maintain robustness to outliers.

![Figure 5. Training Loss](image)

The above graph shows the relationship between the images and their corresponding adversarial and content loss sum at the end of training. An increase in the generator loss decreases the discriminator loss, thus reconstructing the image closer to the original image. While in training phase we can see that the discriminator loss is often less than 0.25 and at some places due to outliers there is a high loss observed at certain images. This can be due to different structural orientations of the face and there is data imbalance with respect to orientations of the face images available. Another reason would be imbalance of images with respect to gender, hairstyle, face attributes and accessories. Thus the facial images are highly challenging to train and the model must be generalized for all orientations irrespective of the above mentioned imbalanced categories. The model has been mainly tested on images of dimension 125x125 and bit depth of 24 in PNG format. Hence, this model can only be evaluated for images with same dimensions and format.

The images generated as a part of inpainted image is then passed to super resolution network and bipolar interpolation is performed on the whole image because the inpainted region and the regular region in the generated image differs, hence the entire image is used to enhance the resolution of the image in order to obtain better results. Figure 6 given below consists of inpainted image obtained as a result of inpainting GAN network. It can be clearly seen that there is an increase in resolution in comparison with the image generated by inpainting network.

![Figure 6. Super-Resolution of Inpainted Image](image)
Table 1. Quantitative results on different methods in CelebA dataset.

| Method                        | Loss Function                                      | PSNR (dB) |
|-------------------------------|----------------------------------------------------|-----------|
| Wasserstein GAN               | L2, global and local discriminator loss            | 27.90 dB  |
| Joint Face Completion Super Resolution GAN | Prior loss, L2 loss(Combination of loss functions) | 27.36 dB  |
| Proposed approach             | L1 loss and MSE loss                               | 28.19 dB  |

The image generated produced better result visually and higher PSNR value than Joint Face Completion Super Resolution GAN and other GAN methods available given in above comparison Table 1.

5. Conclusion
The proposed model for face inpainting and super resolution using GAN provided quality results with test data. Apart from increase in PSNR value to 28.19 dB, this method also involves less time for training and testing phase. While other methods used 14 hours to 3 days for training just for inpainting, our method uses only 12 hours including inpaining and super-resolution. Thus this model is more efficient and provides quality results in less time by using L1 and MSE loss functions. Future implementations of our work include using random masks for inpainting. Though our model uses random masks in training phase, during testing it is necessary to manually set the regions to be inpainted, hence our model will be enhanced to automatically identify the locations that are distorted and abnormal and fill automatically with respect to the neighbouring pixels.

6. References
[1] Kim Muir 2014 Approaches to the reintegration of paint loss: theory and practice in the conservation of easel paintings In Studies in Conservation p 19 – 28
[2] Yi Jiang, Jiajie Xu, Baoqing Yang, Jing Xu and Junwu Zhu 2019 Image Inpainting Based on Generative Adversarial Network in IEEE Access vol. 8 p 22884 – 22892
[3] Dr. Padmavathi S, Dr. Soman K P and Aarthi, R 2018 Image restoration using knowledge from the image in Advances in Intelligent Systems and Computing vol. 177 p 19-25
[4] Ugur Demir and GozdeUnal 2018 Patch-Based Image Inpainting with Generative Adversarial Network In International Conference on Computer Vision and Pattern Recognition, CVPR 2018
[5] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen and Jiayalia 2018 Image Inpainting via Generative Multi-columnConvolutional Neural Network in International Conference on Advances in Neural Information Processing Systems, ICANIPS 2018.
[6] Raymond A. Yeh, Chen Chen, Teck Yian Lim and Alexander G. Schwing 2019 Semantic Image Inpainting with Deep Generative Models in Proceedings of the IEEE Conference Computer Vision and Pattern Recognition, p 5487-5493
[7] Yang Li, Bitao Jiang, Yao Lu and Li Shen 2019 Fine-grained Adversarial Image Inpainting with Super Resolution in International Joint Conference on Neural Networks, IJCNN 2019
[8] Weimin Tan, Bo Yana and Bahetiyaer Bare 2018 Feature Super-Resolution: Make Machine See More Clearly in CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018
[9] Harini S and Aarthi R 2018 A Survey of Deep Convolutional Neural Network Applications in Image Processing International Journal of Pure and Applied Mathematics vol. 118 p 185-189
[10] K. Vikram and Dr. Padmavathi S 2017 *Facial parts detection using Viola Jones algorithm* In International Conference on Advanced Computing and Communication Systems, ICACCS 2017

[11] OpenCV MMLAB, http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.

[12] Han Wang, Wei Wu and Yang Su 2019 *Image Super-Resolution using a Improved Generative Adversarial Network* In IEEE 9th International Conference on Electronics Information and Emergency Communication, ICEIEC 2019

[13] Dr.R.Sudhakar, Mr.K.Raghesh Krishnan and Mr.S.Muthukrishnan 2011 *A Hybrid Approach to Content Based Image Retrieval Using Visual Features and Textual Queries* In Third International Conference on Advanced Computing, ICAC 2011

[14] Zhijiang Li, Haonan Zhu, Liqin Cao, Lei Jiaon,Yanfei Zhong and Ailong Ma 2019 *Face Inpainting via Nested Generative Adversarial Networks* In IEEE Access vol. 7 p 155462 – 155471

[15] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell and Alexei A. Efros 2016 *Context Encoders: Feature Learning by Inpainting* in Computer Vision and Pattern Recognition, CVPR p 2536–2544

[16] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz and Z. Wang et al 2017 *Photo-realistic single image super-resolution using a generative adversarial network* In Computer Vision and Pattern Recognition, CVPR 2017

[17] Zhao Gao, Eran Edirisinghe and Slava Chesnokov 2019 *Image Super-Resolution Using CNN Optimised by Self-Feature Loss* In International Conference on Image Processing, ICIP 2019