Facial Image Reconstruction Using CNN and CNN-based GAN Models

Dev Banerjee, Manikanta Surapathi, Sai Charanya Ponnala

TABLE OF CONTENTS

1. Abstract

Facial image reconstruction is a challenging task that has attracted much attention in recent years due to its numerous real-world applications. In this project, we proposed a deep learning approach for enhancing the quality of low-quality facial images by developing three models: two convolutional neural networks (CNN), and one CNN-based generative adversarial network (GAN).

The CNN-based models take pairs of low-quality and high-quality images as input and learn to reconstruct high-quality images from degraded ones. On the other hand, the GAN model consists of a generator network that generates high-quality images from low-quality ones and a discriminator network that distinguishes between generated and real high-quality images. During training, the generator is trained to generate images that can deceive the discriminator, while the discriminator is trained to differentiate between real and fake images.

We used the Flick-Faces dataset, which consists of 70,000 facial images at a resolution of 1024x1024, to train and evaluate our models. We uniformly degraded 3,000 high-quality images using the same level of compression to create low-quality input and corresponding target pairs.

The experimental results demonstrate that the CNN models can effectively enhance the quality of low-quality facial images, while the GAN model would require larger training time and resources to be trained effectively. However, in other studies, GAN based models have proven to generate high-quality facial images with fine details and textures that were not present in the low-quality input images, making it ideal for challenging facial image reconstruction tasks.

In summary, the work showcases the potential of deep learning models in enhancing the quality of low-quality facial images, and our proposed models have numerous real-world applications in law enforcement, medical imaging, and photography.

2. Introduction

Facial image reconstruction is a critical task in computer vision that has various real-world applications. With the increasing use of images in modern communication and information systems, the ability to restore the quality of degraded facial images has become essential. The availability of high-resolution images is limited, and often, the available images are blurry, pixelated, or of low quality. This issue can be challenging in various fields, including law enforcement, medical imaging, and photography. Hence, developing an effective method for enhancing the quality of low-quality facial images has become a crucial research topic.

Recently, deep learning has shown remarkable success in various computer vision tasks, including image restoration, object detection, and image classification. In particular, convolutional neural networks (CNNs) and generative adversarial networks (GANs) have been widely used for image restoration tasks. In this project, we propose the use of deep learning models, specifically CNN and CNN-based GAN models, to reconstruct high-quality facial images from low-quality images. We trained our models on a dataset of paired low-quality and high-quality facial images to learn the mapping from low-quality to high-quality images.

The pictures shown here are for explanation purposes only.

The methodology used for enhancing the quality of facial images using deep learning models is presented in this report. The dataset used, the experimental setup, and the results obtained are discussed. Finally, the potential applications of the proposed models in various fields are discussed.

3. Motivation

The motivation behind this project is rooted in the growing need to address the critical problem of facial image reconstruction, which has wide-ranging implications for law enforcement, medical imaging, and photography. The importance of high-quality facial images in these fields cannot be overstated. For instance, in law enforcement, CCTV cameras often produce low-quality images that are blurry and pixelated, making it challenging to identify suspects. In medical imaging, low-quality facial images can impede the diagnosis and treatment of medical conditions, while in photography, poor lighting conditions can result in low-quality images that need to be improved.

The degradation of facial images due to pixelation is a common problem that can make it challenging to identify individuals accurately. Often, the faces in images are only a small part of the overall image, and zooming in on the face can result in blurry or pixelated images. This issue is especially prevalent in low-quality images, such as those captured in low light or from CCTV cameras.

Our project aims to address this problem by developing a Deep Neural Network model that can effectively reduce pixelation in degraded facial images. By enhancing the quality of low-quality facial images, our proposed model can provide smoother and less pixelated facial images, which can aid in identifying individuals accurately. Our work has the potential to significantly benefit various real-world applications, including law enforcement, medical imaging, and photography.

The use of deep learning models, particularly CNN and CNN-based GAN models, has shown significant promise in improving the quality of degraded facial images. By learning the mapping from low-quality to high-quality images, these models can generate high-quality reconstructions with fine details and textures that were not present in the low-quality input images.

The development of effective methods for enhancing the quality of low-quality facial images using deep learning models is a critical area of research in the field of computer vision. This project seeks to contribute to this area of research by developing models that can enhance the quality of facial images and provide significant benefits in real-world applications. By using a large dataset of paired low-quality and high-quality facial images and employing advanced deep-learning techniques, we hope to demonstrate the effectiveness of our proposed models and pave the way for future research in this area.

4. Scope of the Project and Utilized Resources

The scope of this project was to develop and evaluate a deep-learning model that can enhance the quality of degraded facial images. To accomplish this, several deep learning architectures were explored, and their effectiveness was assessed. However, the size of training data, depth of the models, and consequently their performance were all limited by the available computing and time resources. Training deep convolutional neural networks for such complex tasks requires significant computational resources and time, which were restricted in availability for this project.

The project was implemented using Python code in the PyTorch framework, a popular open-source machine learning library that provides an efficient and flexible platform for developing deep learning models. PyTorch provides easy-to-use tools for building neural networks and conducting complex operations on tensors. The framework was chosen for its ease of use, scalability, and ability to run on GPUs for faster training.

To cope with the computational demands of the project, the project was executed on Georgia State University's Analytics Research Cluster, a high-performance computing system that provides researchers with significant computational resources. The cluster allowed us to conduct experiments in a parallelized environment, significantly reducing the time required to train the models.

The chosen framework and resources allowed us to best develop and evaluate deep learning models within the frame of given resources.

5. Dataset Description

The dataset used in this project is the Flick-Faces dataset [1], which consists of 70,000 facial images at a resolution of 1024x1024 pixels. The dataset is diverse, including individuals of different ages, genders, and races, making it suitable for training a facial image reconstruction model that can generalize well.

To create pairs of low-quality and high-quality images for training and testing our models, we uniformly degraded 3,000 high-quality images from the Flick-Faces dataset using the same level of compression. The resulting images were pixelated and blurry, simulating real-world scenarios where low-quality images are the only ones available.

The dataset was preprocessed to ensure consistency and compatibility with the deep learning models used in this project. Using this dataset, we trained and evaluated our deep learning models to reconstruct high-quality facial images from low-quality ones, achieving significant visual improvements in image quality.

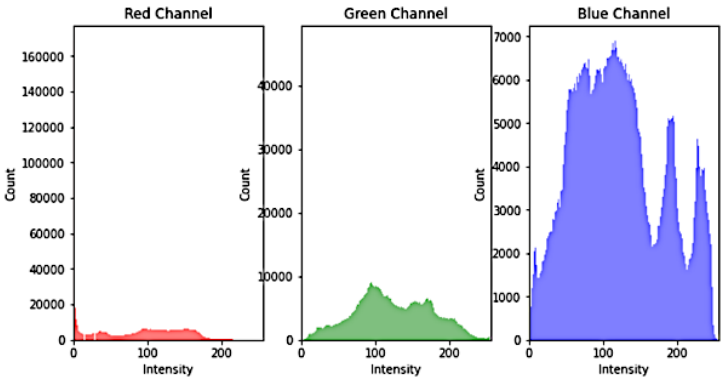
Finally, aside from the 3000 pairs of images used for training, an exclusive set of 20 images were sampled from the original dataset of 70000 images to serve as validation. Similar compression was applied to the original images to obtain low-quality images which served as validation data input, and the high-quality images as validation data output.

Dataset citation-

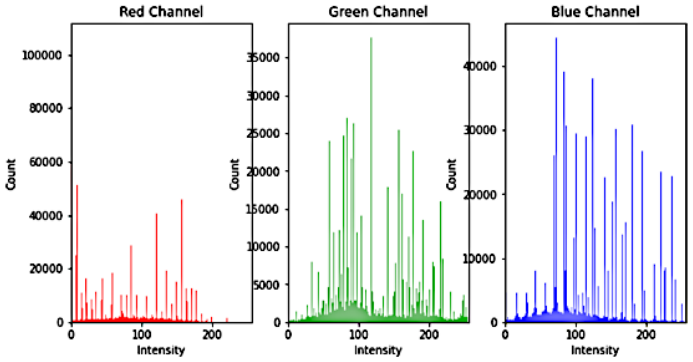
[1] Arnaud, Raphael. (2020). Flickr-Faces-HQ Dataset (FFHQ). [Dataset]. Kaggle. Available from https://www.kaggle.com/arnaud58/flickrfaceshq-dataset-ffhq.

6. Image Degradation

Every high-quality image was uniformly degraded to create corresponding input low-quality images for the model. Degradation involved compressing a range of vivid pixel values into a smaller range, resulting in sharp spikes for certain intensities and gaps for others. This caused blotchy pixels in the degraded image, resulting in the loss of fine details and textures. In contrast, the original images had a wide range of vividly distributed intensity values, with no single value creating sharp spikes. Consider the RGB histograms for an original and corresponding degraded image below-



High Quality Image RGB Intensities



RGB intensities after Compression

Considering the blue channel distribution, it can be observed that several shades of blue are being compressed into one shade during image degradation, which results in large patches of pixels of the same shade of blue. Image degradation led to a drastic reduction in the range of pixel values, causing sharp spikes in the intensity histograms. Additionally, certain intensities were completely lost, resulting in lacunae in the histograms.

This phenomenon is responsible for the pixelation and loss of fine details in the degraded images, which is what was desired in the input images for this task.

The primary goal of our project was to develop a deep learning model that could effectively restore degraded facial images to high-quality images. **Therefore, the**

effectiveness of the model can also be assessed by how well it is able to redistribute the RGB intensities of the input image to mimic the RGB distributions of the target image.

7. Model Selection

Two well-known image manipulation network types were experimented with in order to achieve the objective of reconstructing high-quality images from pixelated, low-quality images: Convolutional Neural Network (CNN) and Generative Adversarial Network (GAN).

1. The CNN model was chosen due to its effectiveness in redistributing RGB intensities by identifying relationships between input and target images. Convolutional layers allow for the identification of local patterns in color features, which can then be used to redistribute RGB intensities to align with the target distribution. However, it should be noted that CNN models have a large number of parameters, which can result in long training times.

2. The GAN model was also explored due to its established ability to generate new images with enhanced details [2]. GAN model consists of a generator and a discriminator, where the generator generates new images that the discriminator evaluates for authenticity compared to high-quality images. During training, the generator is updated to generate images that are difficult for the discriminator to distinguish from high-quality images. Once trained, the generator can generate new images that match the target distribution and can be used to redistribute RGB intensities. However, like CNN models, GANs have a large number of parameters and may require a significant amount of time to train.

Overall, both models types were considered due to their effectiveness in image reconstruction. Two CNN architectures were built to determine the effect of having a deeper network.

[2] Tongxin Wei, Qingbao Li, Zhifeng Chen, Jinjin Liu, "FRGAN: A Blind Face Restoration with Generative Adversarial Networks", *Mathematical Problems in Engineering*, vol. 2021, Article ID 2384435, 14 pages, 2021. https://doi.org/10.1155/2021/2384435

8. Data Pre-processing

Since the objective is to reconstruct high-quality images from low-quality images, there were not many pre-processing steps involved once the degraded images had been generated. This was because preprocessing steps like resizing the images to a smaller size would serve against the project's objective and lead to further degradation of image quality when compared to the bigger image input of size 1024x1024.

Therefore, only the following pre-processing steps were involved-

1. **Normalizing**: The pixel values of the images were normalized to be between 0 and 1. This was done to ensure that the pixel values fall within the range of activation functions used in the models, which helps in faster convergence during training.

2. **Converting to grayscale (GAN Only)**: The GAN model is built of two CNN models and requires double the training time. Therefore, to reduce the computational complexity and training time of the GAN model, both the low-quality input images and high-quality target images were converted to grayscale. This reduces the number of input channels from three (RGB) to one (gray), resulting in faster training times.

9. Model Architectures

For this project, three different architectures were experimented with to determine the most effective model for reconstructing high-quality facial images from pixelated and low-quality input images.

9.1 BasicCNN

The first architecture implemented was called 'BasicCNN', which was a lightweight CNN with three convolutional layers, with a single pixel zero padding on all edges for all convolution layers to preserve some edge data.

```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1       [-1, 64, 1024, 1024]           1,792
              ReLU-2       [-1, 64, 1024, 1024]               0
            Conv2d-3       [-1, 64, 1024, 1024]          36,928
              ReLU-4       [-1, 64, 1024, 1024]               0
            Conv2d-5        [-1, 3, 1024, 1024]           1,731
================================================================
Total params: 40,451
Trainable params: 40,451
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 12.00
Forward/backward pass size (MB): 2072.00
Params size (MB): 0.15
Estimated Total Size (MB): 2084.15
----------------------------------------------------------------
```

Basic CNN

9.2 GAN

The second architecture explored was the Generative Adversarial Network (GAN), which is a more complex model that consists of two CNN networks - the Generator and the Discriminator. Both the generator and discriminator were built with the same architecture as shown below, with a single pixel zero padding on all edges for all convolutional layers. The total number of trainable parameters for this model were 76,290 (38,145 each for generator and discriminator).

```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1       [-1, 64, 1024, 1024]             640
              ReLU-2       [-1, 64, 1024, 1024]               0
            Conv2d-3       [-1, 64, 1024, 1024]          36,928
              ReLU-4       [-1, 64, 1024, 1024]               0
            Conv2d-5        [-1, 1, 1024, 1024]             577
================================================================
Total params: 38,145
Trainable params: 38,145
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 4.00
Forward/backward pass size (MB): 2056.00
Params size (MB): 0.15
Estimated Total Size (MB): 2060.15
----------------------------------------------------------------
```

GAN (Same for Generator and Discriminator)

9.3 BigCNN

The third architecture explored was a deeper version of the BasicCNN, with five convolutional layers. The model was designed to learn more intricate features in facial images. As a resultWhile the BigCNN has a longer training time and more parameters than the BasicCNN, it is more effective in reconstructing facial images. However, this architecture result in the highest number of trainable parameters.

```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1       [-1, 64, 1024, 1024]           1,792
              ReLU-2       [-1, 64, 1024, 1024]               0
            Conv2d-3       [-1, 64, 1024, 1024]          36,928
              ReLU-4       [-1, 64, 1024, 1024]               0
            Conv2d-5      [-1, 128, 1024, 1024]          73,856
              ReLU-6      [-1, 128, 1024, 1024]               0
            Conv2d-7      [-1, 128, 1024, 1024]         147,584
              ReLU-8      [-1, 128, 1024, 1024]               0
            Conv2d-9        [-1, 3, 1024, 1024]           3,459
================================================================
Total params: 263,619
Trainable params: 263,619
Non-trainable params: 0
----------------------------------------------------------------
Input size (MB): 12.00
Forward/backward pass size (MB): 6168.00
Params size (MB): 1.01
Estimated Total Size (MB): 6181.01
----------------------------------------------------------------
```

Big CNN

These three architectures, BasicCNN, GAN, and BigCNN, were explored to determine the most effective model for reconstructing high-quality facial images from low-quality input images. Each model has its strengths and weaknesses, and the model selection process involves balancing model complexity, training time, and model performance.

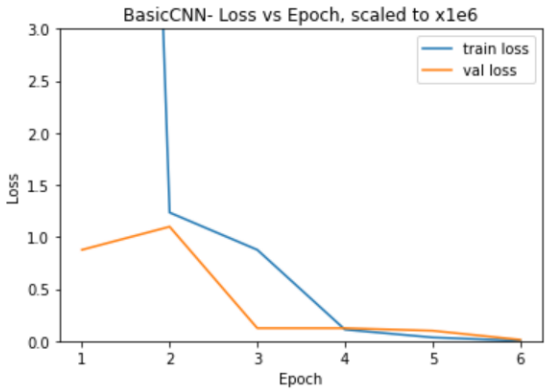
10. Loss Functions

For the **BasicCNN** and **BigCNN** architectures, the **mean squared error (MSE)** loss function was used. This loss function calculates the differences between corresponding pixel intensities of the output and target images. By minimizing the MSE loss, the model is encouraged to generate output images that closely resemble the high-quality target images.

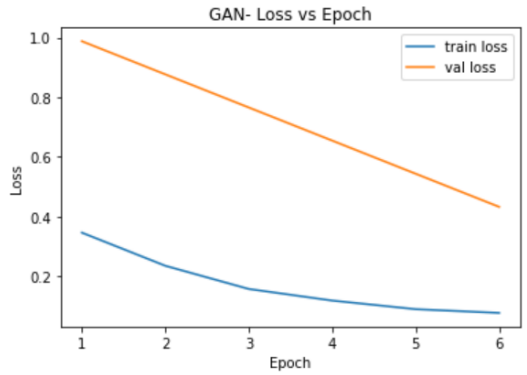
In the case of the **GAN** architecture, the **binary cross-entropy (BCE)** loss function was used. This loss function measures the difference between the generated images and the real high-quality images. The BCE loss penalizes the generator for generating images that are too dissimilar from the high-quality images. By minimizing the BCE loss, the generator is encouraged to produce images that closely match the high-quality images, resulting in better facial image reconstruction. Furthermore, the BCE loss is known to address the problem of vanishing gradients, which can occur when using alternative loss functions, such as mean squared error, for training GANs.

11. Learning Curves

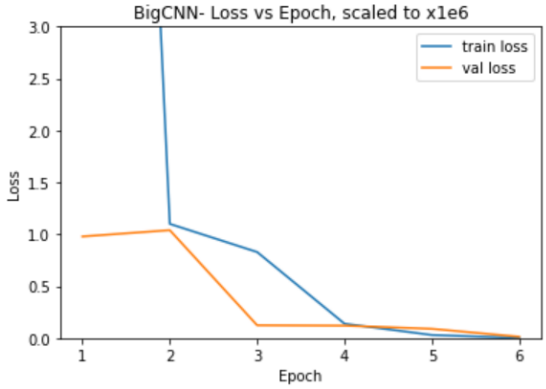
Each of the three models was attempted to train for 30 epochs. The training and validation loss for the first 6 epochs are plotted below-



BasicCNN:
The model was trained effectively for 30 epochs on 3000 images. The model showed effective **learning till the end of the training period, suggesting that a deeper model could be more effective.**



GAN:
The model was trained for only 6 epochs due to massive training times per epoch. However, over the 6 epochs, the model showed improvement in both training and validation loss. This suggested that **increased epochs and training data could improve the model performance.**



BigCNN:
The model was trained effectively for 20 epochs with constantly decreasing validation loss up till the 20th epoch. **The best validation loss at the end of the 20th epoch was significantly lower than the loss of BasicCNN after 30 epochs of training.** However, the total epochs trained were limited by massive training time per epoch.

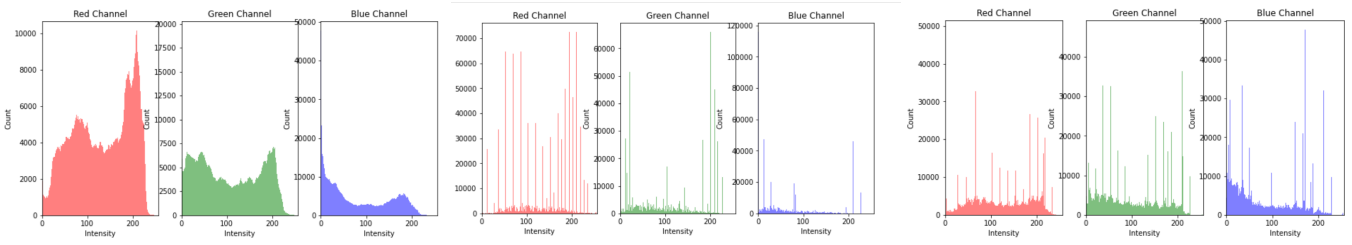
12. Model Results and Comparison

Based on the training and validation losses, it is clear that the BigCNN model is the best performer. However, a sample test image has been used to visualize the results of the three models, and its achieved RGB distributions.

12.1 BasicCNN

Degraded Input (Left) Enhanced Model Output (Right)

The BasicCNN model has clearly performed well in enhancing the visual quality of the degraded image. The distinct pixelated patches of the input image on the right have significantly been reduced and the colors seem to be evened out.
However, there are visible patches of unexpected coloration on the lips, which suggests that the model could be trained further.



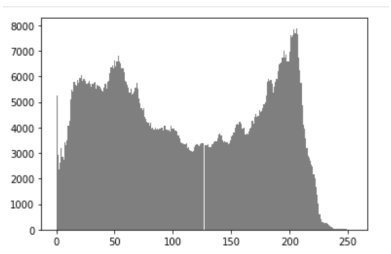
RGB Original (left) **RGB Degraded Input(middle)** **RGB Enhanced Output(right)**

The RGB distributions show that the BasicCNN model has tried to suppress the color spikes of the input image and redistribute them to achieve a distribution inclined towards the original distribution.

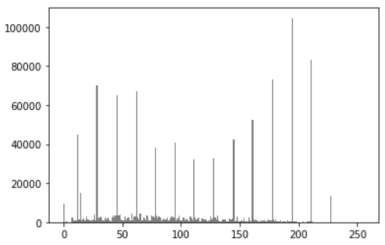
12.2 GAN

Degraded Input (Left) Enhanced Model Output (Right)

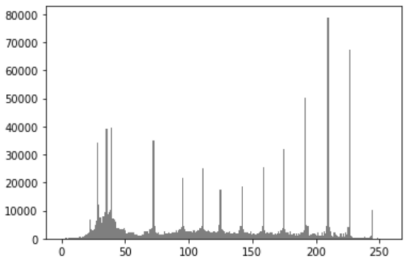
Our implementation of the GAN model could not show any visible improvement in picture quality. However, some improvement was visible in the grayscale intensity distribution.



Intensities Original (left)



Intensities Input(middle)



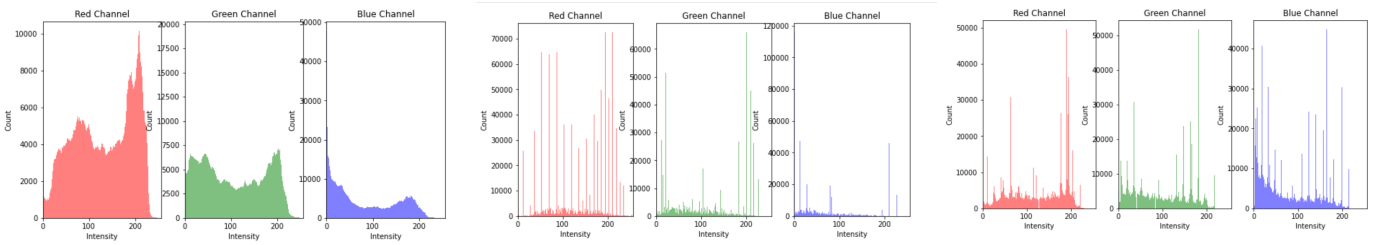
Intensities Output(right)

It can be seen that despite the least number of training epochs, the GAN model has attempted to redistribute the grayscale intensities to mimic the original image. However, this minuscule improvement is not conspicuous as visibly improved quality, which may also be the effect of seeing the image only in a single color channel.

12.3 BigCNN

Degraded Input (Left) Enhanced Model Output (Right)

The BigCNN model seems to have performed at least as well as the BasicCNN model but without the unexpected color patches. Also, the transitions from one shade to the other seem smoother, implying that the achieved RGB distributions are closer to the actual target image.



RGB Original (left) **RGB Degraded Input(middle)** **RGB Enhanced Output(right)**

The RGB distributions of model output are denser in the areas where the target distribution is also dense. However, some intensities are entirely missing from the output distribution, which might be a result of insufficient training data. Overall, the model seems to give a satisfactory improvement in image quality over the input image.

13. Conclusion and Discussion

Out of the three experimented architectures, the **BigCNN architecture outperformed the other two architectures**. The BasicCNN architecture's training curve demanded the model to be made more complicated for this task, which was fulfilled by the BigCNN architecture. The GAN model, on the other hand, could have given better results, however, constraints on computing resources and massive training times with an imposed timeline could not allow for the GAN architecture to be tweaked with.

Through the image degradation process, it was also made clear that **analysis of the image RGB distributions is essential for understanding both the performance of the models and the task of image reconstruction as a whole**.

Further, the training curves of all the models were indicative of **insufficient size of training data**, increasing which will definitely result in improved model performance, and further enhanced images. However, within the scope of this project, the chosen **models performed adequately well, and showed visibly improved image quality of a degraded image** to match the quality of a high-quality, original image. Therefore, in conclusion, **the built CNN and GAN models were successful in the task of Facial Image Reconstruction**, which can be clearly seen in the final comparison set of images with the degraded input image, the enhanced output of the BigCNN model, and the original High Quality image-

Degraded Input (left) **Enhanced BigCNN Output (middle)** **Original (right)**