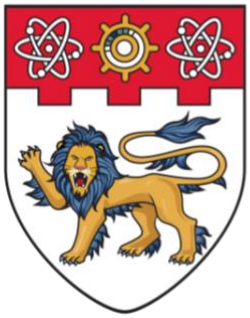


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

SCSE21-0045

Image Quality Improvement on Surveillance Footage using Super-Resolution Techniques to Generate Composite Images

Submitted by: Khalisah binte Faroukh
Matriculation Number: U1822498A

Supervisor: A/P Qian Kemao
Examiner: Asst/P Lana Obraztsova

Submitted in Partial Fulfillment of the Requirements for the Degree of

B.Eng. Computer Science
School of Computer Science and Engineering
Nanyang Technological University

2022

Abstract

Super-resolution is the process of enhancing the resolution of a captured image through the process of upscaling and enhancing a lower-resolution image to construct a higher-resolution image. In recent years, we have witnessed several advancements in the state-of-the-art Deep Learning-based architectures, which also includes remarkable progress in the task of image super-resolution. With many possible interesting applications of super-resolution, we have identified a potential application of super-resolution that could be useful in the field of forensic science, particularly in the area of surveillance.

This project was designed to generate a composite image of the masked face of a suspect captured on a surveillance camera. We have assembled three models that were essentially the baseline model equipped with super-resolution layers for enhancement on the input image and/or on the baseline model's output image. Our models were compared against the baseline model, which is an image inpainting model that does not have any super-resolution layers. All the three proposed models yielded satisfactory results, and one of the proposed models exceeded the performance of the others in terms of visual quality.

This project has successfully demonstrated that the application of super-resolution technique can help in the image quality improvement, which, therefore, will help in generating a more accurate composite picture.

Acknowledgement

First and foremost, I would like to express my gratitude to my final year project supervisor, A/P Qian Kemao for the ever-essential guidance throughout the project duration and for giving me the opportunity and freedom to venture into the area of image super-resolution. I would also like to express my sincere thank you to Ms. Irene Goh who have helped me to acquire the necessary computing resources, equipped with a powerful GPU, which was very essential for me to work on my FYP.

Of course, this project could not have been completed without the constant motivation and support from my family and friends.

Table of Content

List of Tables

List of Figures

1. Introduction

1.1. Background

Image super-resolution (SR) is a computer vision technique that aims to reconstruct a higher-resolution (HR) image from a given lower-resolution (LR) image counterpart. It is an important class of image processing techniques as it has a wide variety of real-world applications that could deliver real benefits. Without SR, the components necessary to capture high-quality images are too expensive to set up or implement [1]. Hence, the use of image processing algorithms presents an alternative that is relatively inexpensive to implement for situations where a high-quality imaging system cannot be incorporated.

With this, alongside the advancement in deep learning techniques in recent years, many researchers have extensively explored SR models to propose techniques to help solve this problem. Deep learning approaches have been fairly successful in solving the issue of image resolution and have often achieved state-of-the-art performance on various benchmarks of SR. Some deep learning methods that have been proposed and applied to deal with SR range from the early Convolutional Neural Network (CNN) based methods, to recent approaches using the Generative Adversarial Networks (GAN). The different SR algorithms generally differ from each other in the aspect of the type of network architectures used, the loss function, metrics as well as the different types of learning principles and strategies, etc. [2]. SR has been progressing and upgrading which opens too many new possibilities of the applications of SR in commercial or research purposes, and one probable application is in the field of forensics science.

In the field of forensic science, digital images or videos can prove to be very beneficial in tactical investigations or evidence used in casework. These footages that are obtained by surveillance cameras can be used by law enforcement to identify suspects or be used as evidence in criminal cases [3]. However, oftentimes, these images or videos are of a lower resolution, usually due to the physical limitations of the device that captured such images, or the post-processing of the digital image. As mentioned previously, minimizing these limitations often requires distinctive components to capture high-quality images, but that is typically too expensive to incorporate. On top of that, a surveillance camera is generally made for bulk collection, and 99.9% of the captured footage is purposeless. As they are

constantly engaged and have a high volume of footage, they are often compressed for storage space, which diminishes quality [4]. Thus, being able to obtain a HR image from a LR image is desirable for such purposes.

1.2. Objective and Scope

To help solve crimes, law enforcement has used facial compositing that is either sketched by hand or digitally created by facial composite software. While composite images are normally sketched based on eye witness' description, motivated by the same idea, this project intends to unmask a masked face that was captured from a surveillance camera to generate a composite image of the face. This project was established with the assumption that criminals would put on face masks to hide their identity from a surveillance camera. Hence, this project aimed to test if the application of super-resolution technique can help in the image quality improvement on lower resolution images that could potentially be retrieved from surveillance footage to generate composite images. The project involves the use of super-resolution to enhance the image resolution, as well as image inpainting to fill in missing parts of the face that was covered by the face mask.

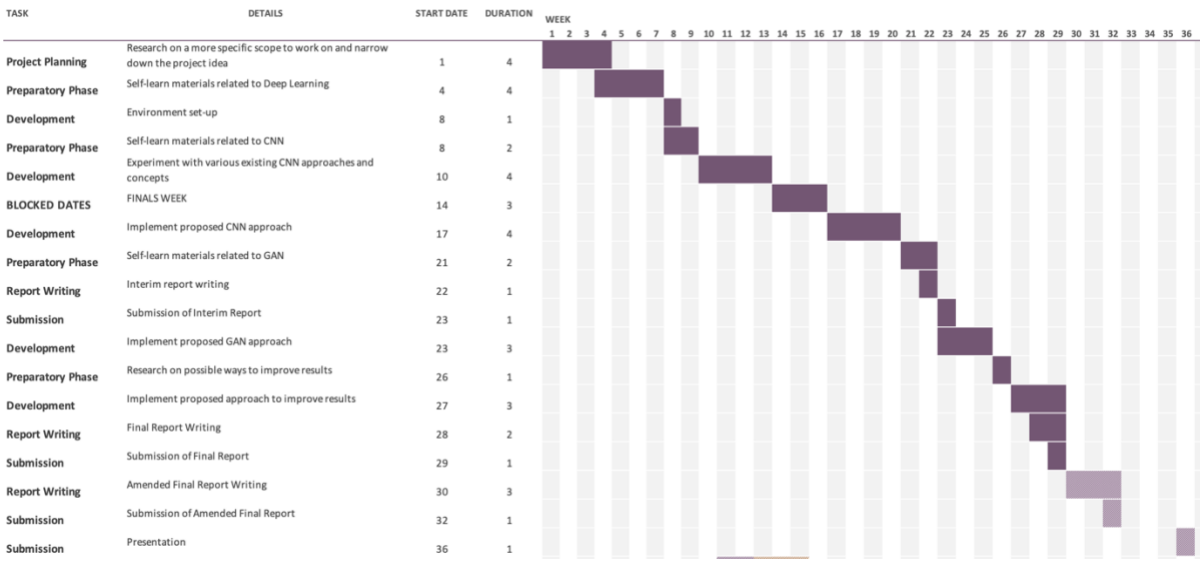
The scope of the project includes building a suitable dataset, assembling the models and then training those models for 2x and 4x upscaling factors on the constructed dataset, unmasking the masked face on the image, and finally, evaluating the results of whether the various super-resolution technique will improve the accuracy of the composite image resemblance to the ground truth.

1.3. Contribution

This project contribution is as follows:

1) Assembled model 1 that uses FSRCNN as its super-resolution technique, conjoint with image inpainting, which produced images with artifacts and edge trails.
2) Assembled model 2 that uses ESRGAN as its super-resolution technique, conjoint with image inpainting, which produced images with more textured details and reduced artifacts.
3) Assembled model 3 that uses ESRGAN as its base super-resolution technique, conjoint with image inpainting, and GFPGAN to enhance the post-processed image which had produced images that are substantially more natural and believable.

1.4. Project Schedule



TASK	DETAILS	START DATE	DURATION
Project Planning	Research on a more specific scope to work on and narrow down the project idea	1	4
Preparatory Phase	Self-learn materials related to Deep Learning	4	4
Development	Environment set-up	8	1
Preparatory Phase	Self-learn materials related to CNN	8	2
Development	Experiment with various existing CNN approaches and concepts	10	4
BLOCKED DATES	FINALS WEEK	14	3
Development	Implement proposed CNN approach	17	4
Preparatory Phase	Self-learn materials related to GAN	21	2
Report Writing	Interim report writing	22	1
Submission	Submission of Interim Report	23	1
Development	Implement proposed GAN approach	23	3
Preparatory Phase	Research on possible ways to improve results	26	1
Development	Implement proposed approach to improve results	27	3
Report Writing	Final Report Writing	28	2
Submission	Submission of Final Report	29	1
Report Writing	Amended Final Report Writing	30	3
Submission	Submission of Amended Final Report	32	1
Submission	Presentation	36	1

Figure 1-1: Project schedule during AY21/22. Week 1-13 are in the first semester. Week 17-21 are during winter break. Week 22-36 are in the second semester.

1.5. Organisation

The report is organised in the following manner:

1) Section 1 introduces our project, along with the objectives, contribution, and project schedule.
2) Section 2 gives us a literature review of the concepts that we will be utilizing at the base of our project
3) Section 3 presents the evaluation metrics used to assess our model performance, the loss functions used for the different models, the steps taken to prepare the dataset, as well as the design and architecture of our models used in the experiments.
4) Section 4 contains the details of our experiment, the results of the proposed methods, and our analysis of the results obtained.
5) Section 5 presents our recommendations for future work.
6) Section 6 contains the conclusion of this report.

2. Review of Theory and Previous Work

2.1. Super-Resolution Networks

2.1.1. CNN Based Approaches

While there has been much work on image super-resolution using pixel-wise interpolation before the advent of convolutional neural networks (CNNs), CNN-based approaches have shown excellent performance with high potential. The first breakthrough in super-resolution with a fully convolutional neural network was SRCNN [5] by Dong et al. which learns an end-to-end mapping from a bicubic interpolation to upscale a low-resolution image to a high-resolution image. SRCNN is arranged into a three-stage process:

1) Patch Extraction and Representation: This operation is used to extract dense patches from the low-resolution image, and projects them into high-dimension LR feature space
2) Non-Linear Mapping: This operation non-linearly maps the high-dimension LR feature space into another high-dimension HR feature space which results in a HR patch
3) Reconstruction: In this final layer, the generated HR patch from the previous operation is aggregated to produce the reconstructed super-resolution image output

Due to its simple network architecture and exceptional restoration quality, SRCNN has drawn attention, however, the processing speed on large images is unsatisfactory. Hence, the researchers proposed another model, which is FSRCNN [6]. FSRCNN is arranged into a five-stage process:

1) Feature Extraction: This operation is used to extract feature maps directly from the low-resolution image
2) Shrinking: This operation reduces the dimension of the feature vector by using a small number of filters
3) Non-Linear Mapping: This operation non-linearly maps the feature maps that represent the LR patches to HR ones. In this step, multiple mapping layers are used with a smaller filter size as opposed to the single-wide convolutional filter used in SRCNN. The number of mapping layers used in this step affects the mapping accuracy and complexity
4) Expanding: This operation performs the inverse of the shrinking layer to increase the dimension of the feature vector

5) Deconvolution: In this final layer, the inverse of convolution is performed to produce the reconstructed super-resolution image output from the HR features

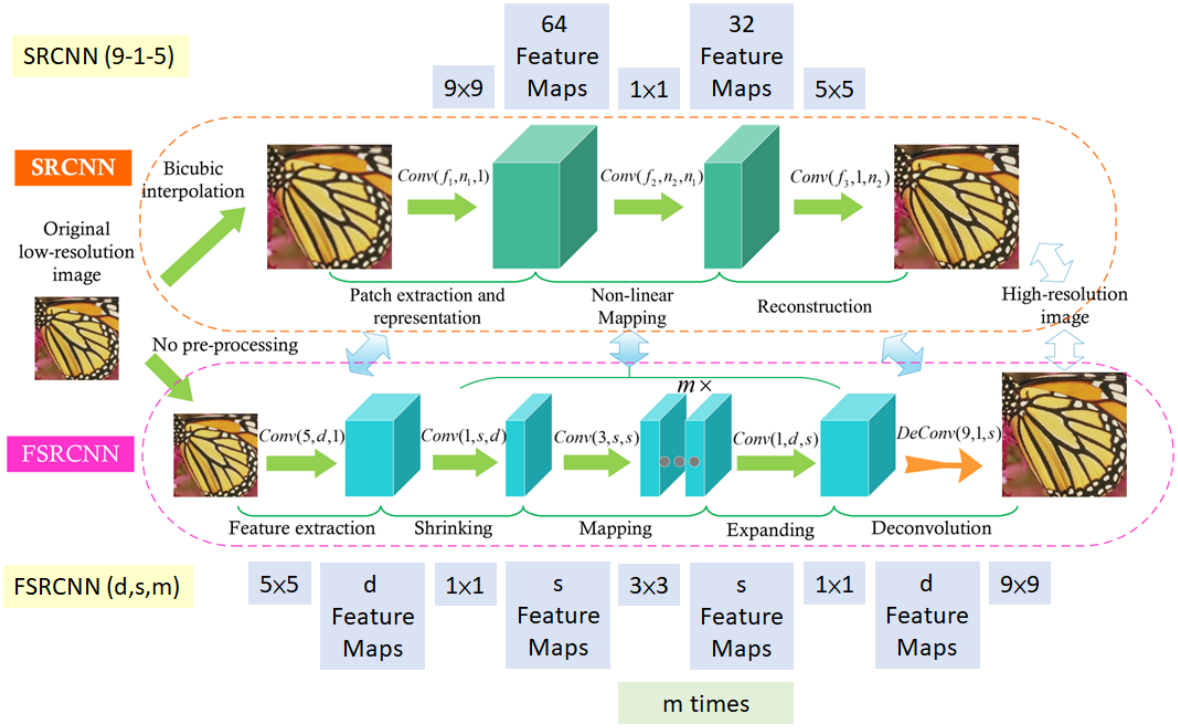


Figure 2-1: Comparison of SRCNN and FSRCNN network structure (obtained from [6])

From this, we could see that the proposed FSRCNN differs from SRCNN in three different aspects. Firstly, we can see that there is no pre-processing or upsampling done in the first stage. Feature extraction had taken place in the low-resolution image space without bicubic interpolation. Secondly, the shrinking stage decreases the number of channels with a (1x1) convolution, thus there is lesser computation and memory. Thirdly, instead of having a single wide convolutional filter, multiple (3x3) mapping layers are used. Lastly, upsampling is done only in the final stage using a (9x9) deconvolutional filter. Hence, this has given FSRCNN a better performance.

2.1.2. GAN Based Approaches

Despite the breakthrough in the speed and precision of a super-resolution image with convolutional neural networks, the output often generates blurry images that lack high-frequency details at large upscaling factors. Hence, Christian et al. [7] had proposed SRGAN

architecture which solves this problem to generate high-quality, state-of-the-art images. In this proposed model, they had utilized a deep residual network (ResNet) with skip connections and introduced a perceptual loss function instead of using the Mean Squared Error loss. This has helped in recovering the finer details of the image and removing the over-smoothing effects of MSE loss.

Similar to GAN architectures, SRGAN also contains two important network architectures:

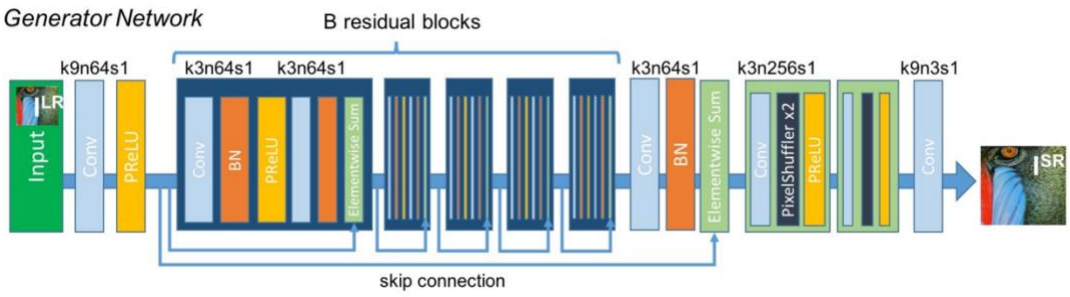


Figure 2-2: Architecture of Generator network (obtained from [7])

1) Generator Network: This architecture utilizes the SRRESNET fully convolutional model. The model begins by passing the low-resolution image input through a (9x9) convolutional layer and 64 feature maps, followed by a Parametric ReLU activation function. The next layer, as illustrated in Figure 2-2, is the core of SRGAN which uses a bunch of residual blocks. Each residual block contains a (3x3) convolutional layer and 64 feature maps, followed by a batch normalization layer, a Parametric ReLU activation function, and an elementwise sum method. Once the residual blocks are formed, the rest of the model is constructed, as illustrated in Figure 2-2. The input image resolution is multiplied with two trained sub-pixel convolutional layers.

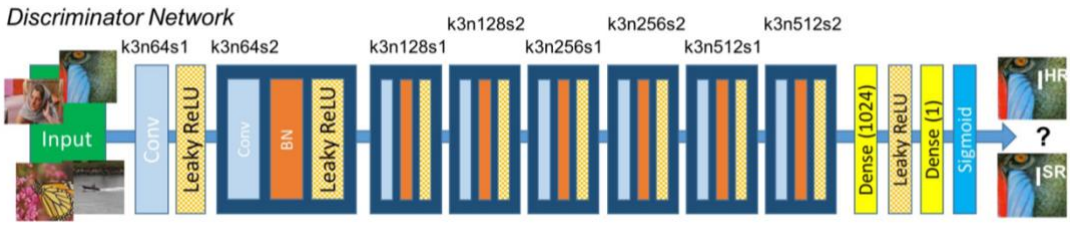


Figure 2-3: Architecture of Discriminator network (obtained from [7])

2) Discriminator Network: This architecture serves as an image classifier that distinguishes between the real HR image and the generated SR image. The model begins with an initial convolutional layer, followed by a Leaky ReLU activation function. The next layer is a bunch of repeating blocks of convolutional layers with an increasing number by a factor of 2 of a 3x3 filter kernels from 64 to 512. It is followed by a batch normalization layer and the Leaky ReLU activation function. Once these repetitive blocks are formed, we have two dense layers and a final sigmoid activation function for performing the classification action.

Xintao et al. [8] proposed ESRGAN to improve the SRGAN architecture by modifying three aspects:

1) Residual-in-Residual Dense Block (RRDB) without batch normalization was introduced. Researchers have observed that batch normalization tends to bring artifacts when the network is deeper and trained under a GAN framework, hence, they have removed the BN layer to improve the generalization ability and reduce computational costs.

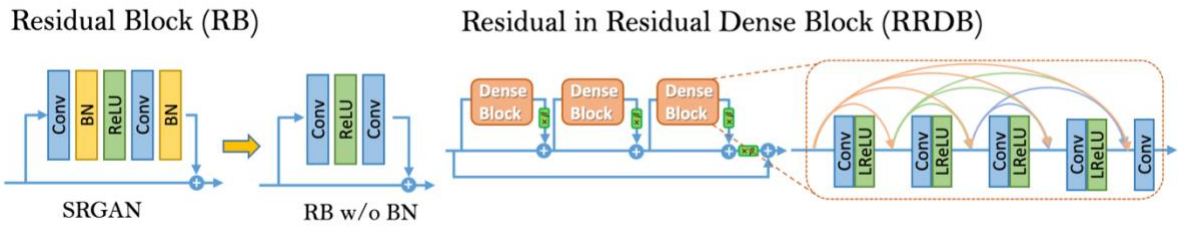


Figure 2-4: Comparison of RB and RRDB (obtained from [8])

On top of that, they have proposed to use dense block instead of residual block, which results in the network capacity becoming higher and have a deeper network, hence, boosting the performance of the model.

2) Improved the discriminator network using Relativistic average GAN (RaGAN). This discriminator estimates the probability that a real image is relatively more realistic than a fake one as opposed to SRGAN standard discriminator which only classifies the input image if it is real or fake.

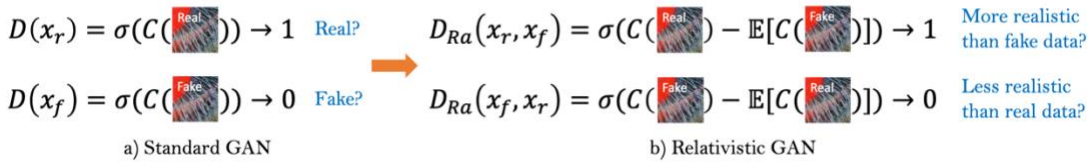


Figure 2-5: Comparison of standard discriminator and relativistic discriminator (obtained from [8])

3) Improved perceptual loss function by using the features before activation function instead of after as in SRGAN. This is to overcome two drawbacks of the original design. First, as seen in Figure 2-6, the activated features are very sparse, especially on very deep networks. Hence, this causes weak control which in turn causes inferior performance. Second, using the features after activation causes inconsistencies in the reconstructed brightness compared to ground truth.

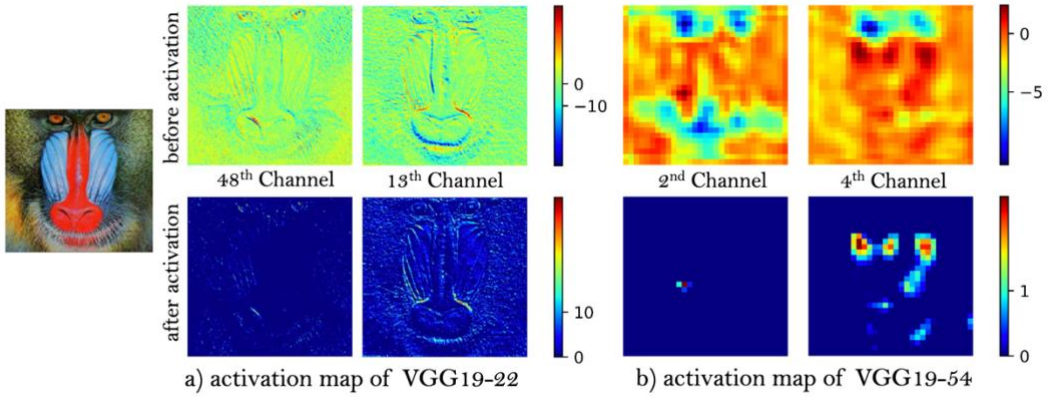


Figure 2-6: Comparison of feature maps before and after activation function (obtained from [8])

GFPGAN [9] was also proposed for blind face restoration that aims to restore high-quality faces from low-quality counterparts that were caused by any kind of degradations, including low-resolution, blur, noise, compression artifacts, etc.

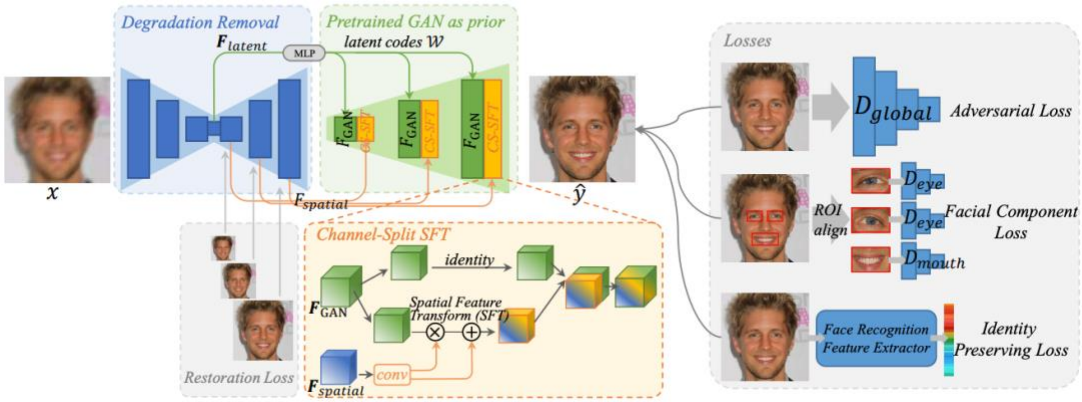


Figure 2-7: GFPGAN framework process (obtained from [9])

As illustrated in Figure 2-7, GFPGAN comprises of the following process:

1) Degradation Removal Module: Using the Vanilla U-Net structure, this module was designed to take a degraded photo and explicitly remove any degradation and at the same time, extract any hidden 'clean' features.

$$\mathrm{F}_{latent,}\mathrm{F}_{spatial} = \mathrm{U} - \mathrm{Net}(x)$$

The following two features are extracted:

a. F_{latent}: The latent feature which maps the input image to the closest latent code in StyleGAN2 [10]
b. $\mathrm{F}_{spatial}$: The multi-resolution spatial feature used to modify the StyleGAN2 features.

2) Generative Facial Prior (GFP) and Latent Code Mapping: We had leveraged a pre-trained StyleGAN2 to act as a generative facial prior to supply varied facial details. Given the latent features F_{latent} of the input image, they are mapped into style vectors by several multi-layer perceptron layers. These vectors are then passed through each convolutional

layer in the pre-trained StyleGAN2 to produce GAN features, which are then further modified by the spatial features $F_{spatial}$.

3) Channel-Split Spatial Feature Transform. This allows the spatial features $F_{spatial}$ to generate a pair of affine transform parameters that can be used to scale and shift the GAN features. For a better balance of realness and accuracy, the spatial alteration is only done on part of the GAN features, whereas others are left to pass through unchanged if the model does not find a need to change them.

4) Finally, the reconstruction loss, adversarial loss, facial component loss, and identity preserving loss of the generated images were implemented for further refinement on the images until training is completed.

2.2. Image Inpainting

Over the last few years, deep learning has been growing very rapidly and has been used in many fields, one of them being image completion. Image completion, alternatively known as image inpainting, is the task of filling in missing regions of an image. It can be used to restore the holes left after removing unwanted objects from an image.

There are many different approaches used in image inpainting, and one of the approaches uses partial convolution. This approach has a U-Net architecture which replaces all convolutional layers to partial convolutional layers. In the decoding stage, nearest neighbour up-sampling is employed. U-Net has proved to be very successful on tasks with output that require a different amount of spatial resolution or on output that has a similar size as its input counterpart. This makes them a very powerful image generation tool.

They implement skip connections which concatenates the identity from the deconvolutional blocks to the corresponding upsampling blocks. This can be seen in Figure 2-8, where the grey arrow illustrates this operation.

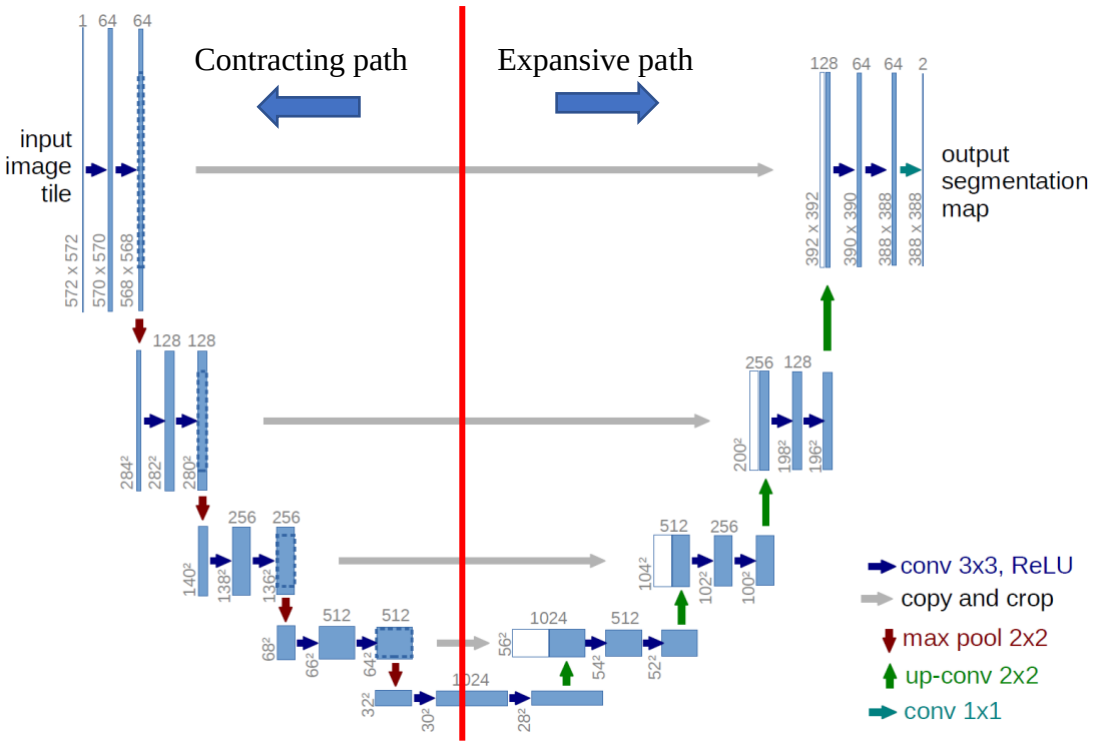


Figure 2-8: U-Net architecture (obtained from [11])

Figure 2-8 shows the U-Net architecture example for a 32x32 pixel image in the lowest resolution. Each blue box represents the multichannel feature maps. The number of channels is denoted on top of the box whereas the x-y-size is provided at the lower-left edge of the box. The white boxes represent the copied feature maps, and the arrows denote the different operations.

The goal of our image inpainting model is to virtually remove the mask from a person's face. Hence, skip connections with U-Net helps to preserve the parts of the input image that will be copied onto the output, whereas the Encoder-Decoder part of the U-Net detects the mask on the face, and replaces it with the predicted nose and mouth underneath the mask.

The image inpainting model was trained with U-Net architecture using the ADAM optimizer and SSIM loss function. The SSIM metric was selected as it gives a numeric value expression of how much the two images resemble each other. To train this model, the dataset was split into testing (20,000 images) and training (the rest of the dataset).

3. Methodology

3.1. Evaluation Metrics

3.1.1. Peak Signal to Noise Ratio (PSNR)

PSNR is a quality measurement metric used to measure the peak error between an original image to a constructed SR image. Having a higher PSNR value implies that the quality of the SR image that has been reconstructed matches the original image much better, and it has a better reconstructive algorithm. PSNR is computed as follows, where R represents the maximum fluctuation in the input data image:

$$PSNR = 10\ log_{10}(\frac{R^2}{MSE})$$

The MSE represents the cumulative squared error between the SR image and the original image.

3.1.2. Structural Similarity Index Measure (SSIM)

SSIM is a perceptual metric used to quantify the similarity between an original image to a constructed SR image. It uses the properties of the human visual perception system under the assumption that it identifies structural information from a scene. Hence, SSIM identifies the differences between the information extracted from the original reference image and the SR image scene. This metric inspects the change in an image structure and extracts the pixel interdependencies, luminance, and contrast features from an image.

SSIM values are in the range of [0, 1], where the higher the value, the more similar the two images are. The luminance is measured by averaging over all pixel values. The contrast is measured by taking the standard deviation of all pixel values. Finally, the structure features are measured by dividing the input image by its standard deviation to get a unit standard deviation. The figure below illustrates the arrangement and flow of SSIM.

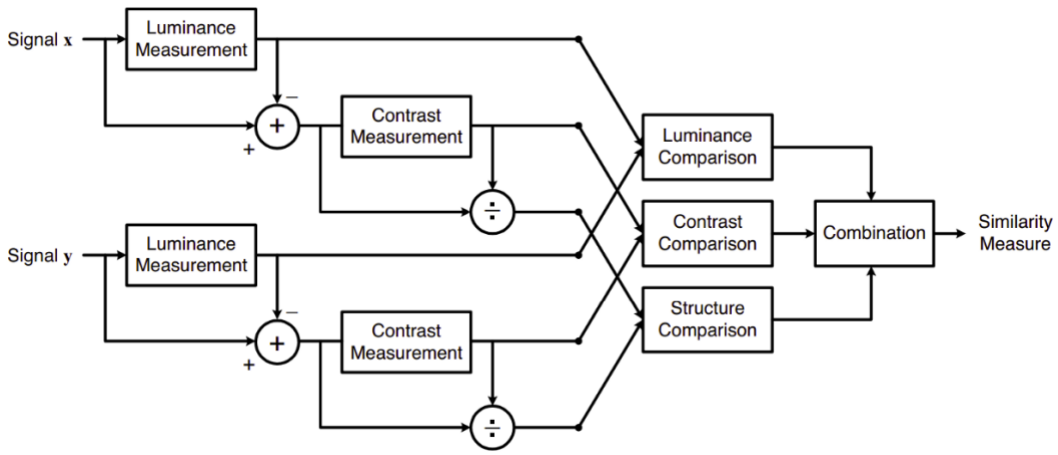


Figure 3-1: The SSIM system. Signal x represents the original reference image, whereas signal y represents the SR image (obtained from [12])

3.1.3. Naturalness Image Quality Evaluator (NIQE)

NIQE is a blind image quality assessment metric that is based on measurable deviations from statistical regularities observed in natural images [13]. Unlike other blind image quality assessments, NIQE does not require training on human-rated distorted images nor need exposure to distorted images.

This metric does not need any ground truth images to be used as a comparison for the reconstructed image. Instead, NIQE compares the image to a default model computed from images of natural scenes. Hence, the value obtained is the degree to which the reconstruction looks like a valid natural image. Having a lower NIQE score implies that the quality of the SR image that has been reconstructed is better in terms of visual quality.

3.2. Loss Functions

3.2.1. Mean Squared Error (MSE) (see FSRCNN)

MSE is a common metric used as an objective measure of distortions. Its value represents the average difference of the pixels all over the image. Having a higher value implies that there is a greater difference between the original ground truth image and the SR image. MSE is computed as a mean of the sum of squared difference, where O represents the original image, and S is the SR image:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)^2$$

3.2.2. VGG Loss (SRGAN)

VGG Loss is a type of perceptual loss function, which is a combination of content loss and adversarial loss. Perceptual loss is a term in the loss function that favours outputs that is close to a human's perceptual level. This loss function helps in creating a more natural and realistic image output. VGG loss can be interpreted as follow:

$$l^{SR} = \underbrace{\underbrace{l^{SR}_{\mathrm{X}}}_{\text{content loss}} + \underbrace{10^{-3} l^{SR}_{Gen}}_{\text{adversarial loss}}}_{\text{perceptual loss (for VGG based content losses)}}$$

Figure 3-2: VGG Loss formula (obtained from [7])

In this loss function, the high-level features of the ground truth images and the generated images from the generator are extracted, then the Euclidean distance between the two feature representations is calculated as content loss [14]. Unlike MSE, this loss function is only concerned about the quality improvement of the images over the pixel-by-pixel comparison of two images.

3.3. Dataset Preparation

In this project, we had experimented with two different datasets. For fast training, we had initially started our experiment using the Labeled Faces in the Wild dataset from the University of Massachusetts [15] which contained over 5,000 images with a focus on human faces. However, we noticed that this dataset is not high-resolution and did not give us a good final image reconstruction result as seen in Figure 3-3, therefore, for our final experiments, the CelebA-HQ [16] dataset was used to carry out the experiments.

Figure 3-3: Results obtained from Labeled Faces in the Wild dataset

The CelebA-HQ dataset contains high-quality images and has above 30,000 sample images. However, as this dataset does not contain any masked images, nor are there any available datasets that contain a high-resolution image of masked faces, a substitute was required. Hence, the MaskTheFace [17] tool was used to synthetically apply masks to the faces and convert the CelebA-HQ dataset to simulate a masked face dataset.

MaskTheFace tool is a computer vision script that uses the dlib based face landmarks detector to identify face tilts and the key features of a face to synthetically warp the masks to the face [17].

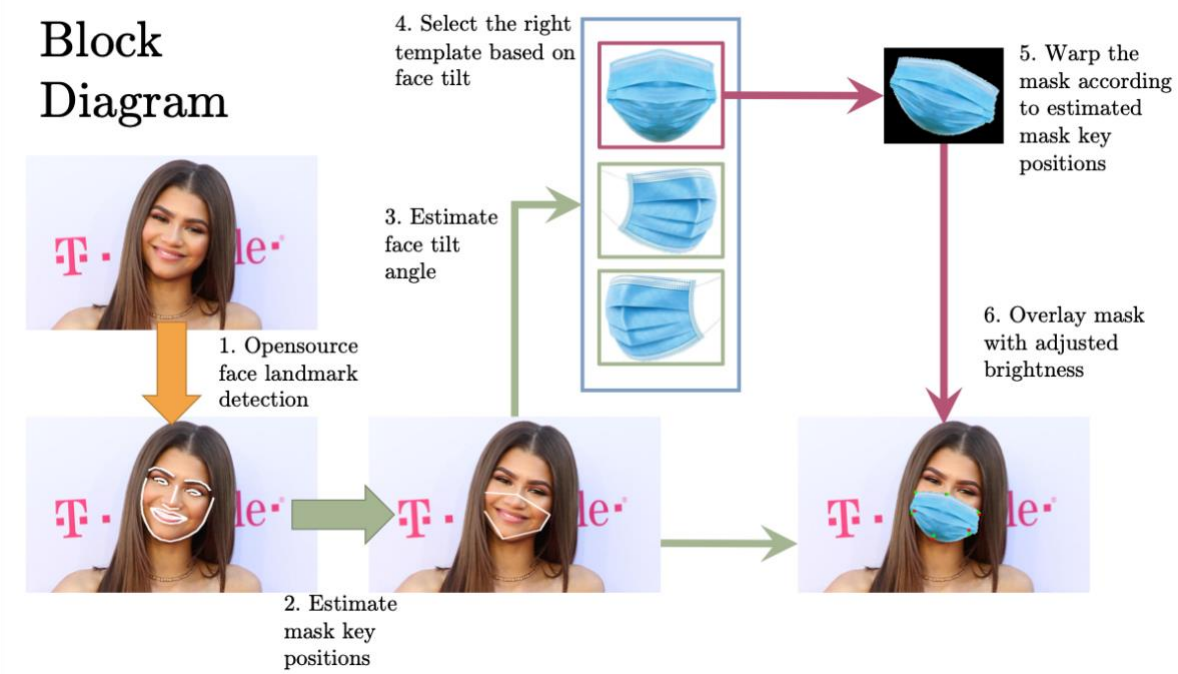


Figure 3-4: MaskTheFace program flow (obtained from [17])

The tool supports 5 types of masks: Surgical, N95, KN95, Cloth, and Gas, and has 24 patterns that can be applied to the mask to create more variations, as well as allow flexibility in the variation of colours and intensity.

The type of mask to apply onto each face was selected at random, and both the images with and without the mask were saved into dedicated input and output folders respectively.

We split our dataset into 8:2 for training and testing. To evaluate our model's accuracy and for our experiments, we had extracted a few images from the testing sample which has never been seen by our network's model.

3.4. Design and Architecture of Models

Our main goal is to test if our assembled model can upscale a LR image to a SR image, while also detecting if alongside the application of a SR model, as the quality improves gradually, a more precise composite image can be generated.

Intuitively true in theory, having a higher PSNR or SSIM implies that the model should be able to generate a composite image with greater precision. Hence, we will be comparing our

evaluation results using the evaluation metrics elaborated in Section 3.1 on the baseline model and against all three of our models to prove our hypothesis that an SR model would improve the image quality so that a more accurate composite image can be generated.

Generally, all three models were built by introducing a super-resolution model to the bottom of the baseline model. All the models have the flexibility of learning and testing across the different upscaling factors. In our experiment, we had tested 2x and 4x upscaling factors on the input images.

3.4.1. Baseline Model

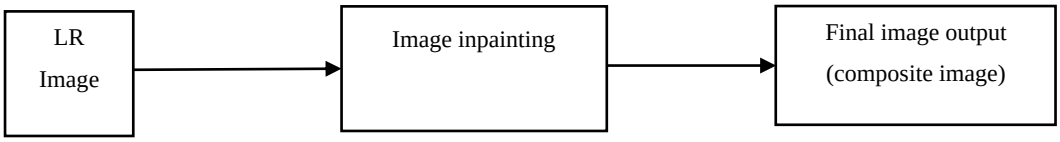


Figure 3-5: Architecture of Baseline Model

Our baseline model is an image inpainting model which was built with ResNet blocks combined with a U-Net architecture. Figure 3-5 illustrates the overall architecture of our baseline model.

The input image is a LR image that contains a masked face, which will be passed into the baseline model, where we would get an unmasked image of the face as the output.

3.4.2. Assembled Model 1

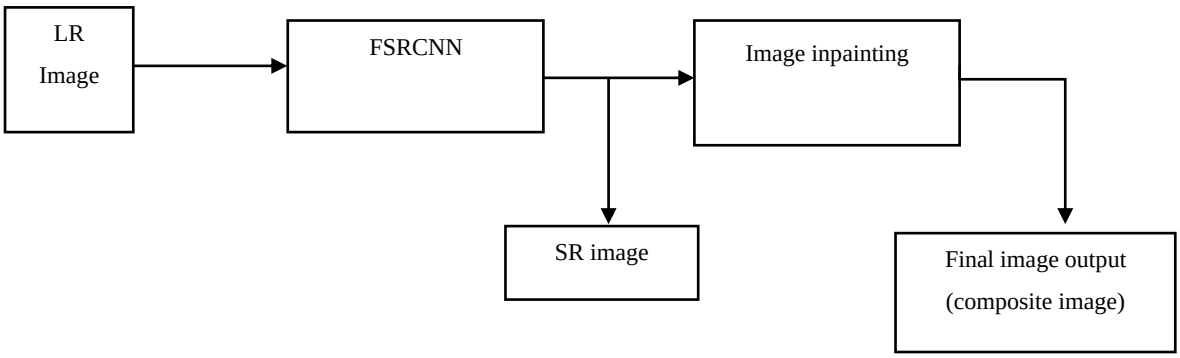


Figure 3-6: Architecture of Assembled Model 1

In our first assembled model, we had introduced FSRCNN to the bottom of the baseline model with the purpose of enhancing the original input image. We have decided the parameters of the model, that is, the filter and kernel sizes, based on the best settings that were presented by Dong et al. on the Set5 dataset [6]. As seen in our architecture in Figure 3-6, the SR image output from the FSRCNN model was passed into our baseline model for composite image generation and is also passed out as a global output.

The loss function of the SR model was computed with the MSE of the SR image and the HR image, and the image generation loss function was computed with SSIM of the original image and the generated image.

3.4.3. Assembled Model 2

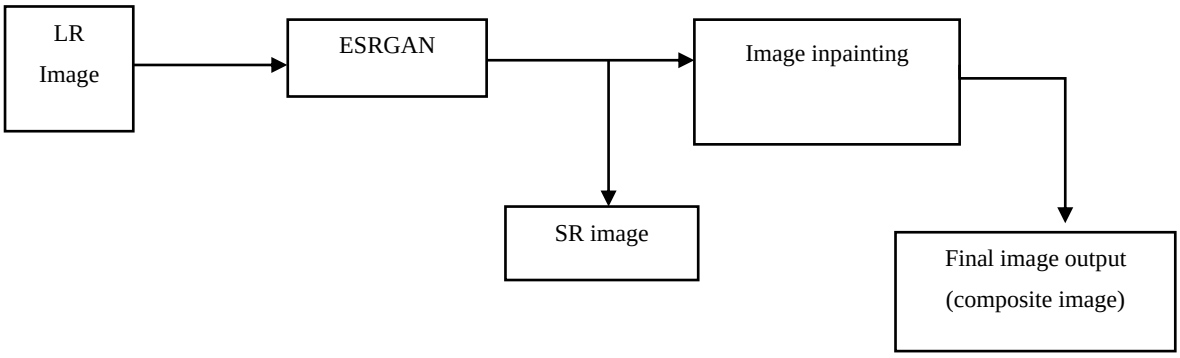


Figure 3-7: Architecture of Assembled Model 2

In our second assembled model, we had introduced ESRGAN to the bottom of the baseline model to enhance the input image. The main difference in the approach between this model and the previous assembled model 1 is the usage of RRDB that was said to deploy a deeper and more complex network structure, thus, boosting performance.

As seen in our architecture in Figure 3-7, the SR image output from the ESRGAN model was passed into our baseline model for composite image generation and is also passed out as a global output.

The loss function of the SR model was computed with the VGG loss of the SR image and the HR image, and the image generation loss function was computed with SSIM of the original image and the generated image.

3.4.4. Assembled Model 3

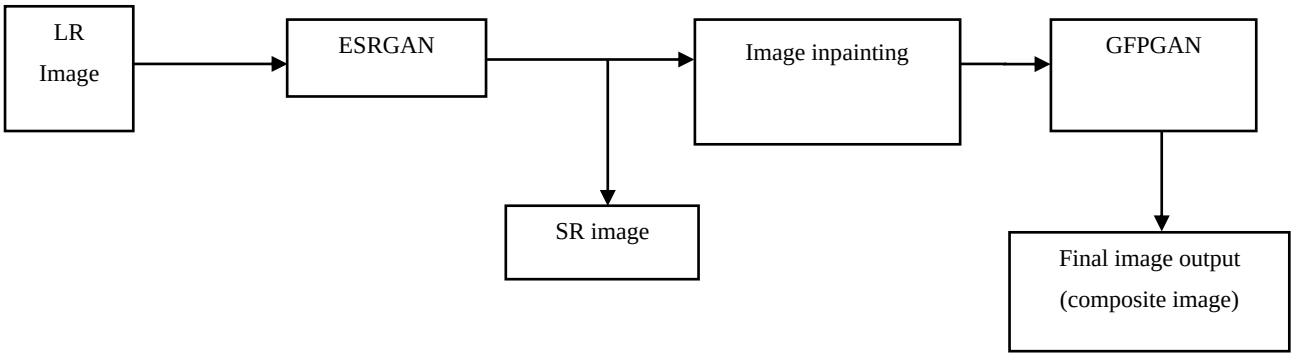


Figure 3-8: Architecture of Assembled Model 3

In our third assembled model, we had introduced GFPGAN at the final stage before producing our final output for the purpose of enhancing the facial image produced by the image inpainting model. GFPGAN was designed to enhance the image quality that contains human faces which are damaged, aged, or otherwise low resolution.

As seen in our architecture in Figure 3-8, the SR image output from the ESRGAN model was passed into our baseline model for composite image generation which is then passed into the GFPGAN model for further image refinement.

4. Experiments and Analysis

4.1. Comparison of Baseline Model with Assembled Model 1

4.1.1. Experiment

The baseline model was trained using the prepared CelebA-HQ dataset as elaborated in Section 3.3. To test the effectiveness of the super-resolution model, a low-resolution image was required. Hence, we had downscaled the HR testing images using bicubic interpolation according to the selected scaling factors. The original image was also preserved so that each image have a ground truth HR image and a LR image to perform quality performance assessment.

The super-resolution model training parameters were kept the same as their original settings for the performance comparison so that its performance will not be undermined. The model

was trained for different upscaling factors. The architecture of both the baseline model and the assembled model is elaborated in sections 3.4.1 and 3.4.2.

To test if super-resolution helps with generating a composite image with greater precision, a quantitative evaluation was done by comparing the infilled output to the original image and computing 2 reconstruction metrics: SSIM and PSNR. Both of these metrics are commonly used metrics for image quality assessment. A detailed description of these metrics can be found in sections 3.1.1 and 3.1.2.

A qualitative evaluation was also done by assessing the image output and firstly, ensuring that there are minimal visible image artifacts, secondly, it looks natural to the human eye, and finally, it is not pixelated.

4.1.2. Results

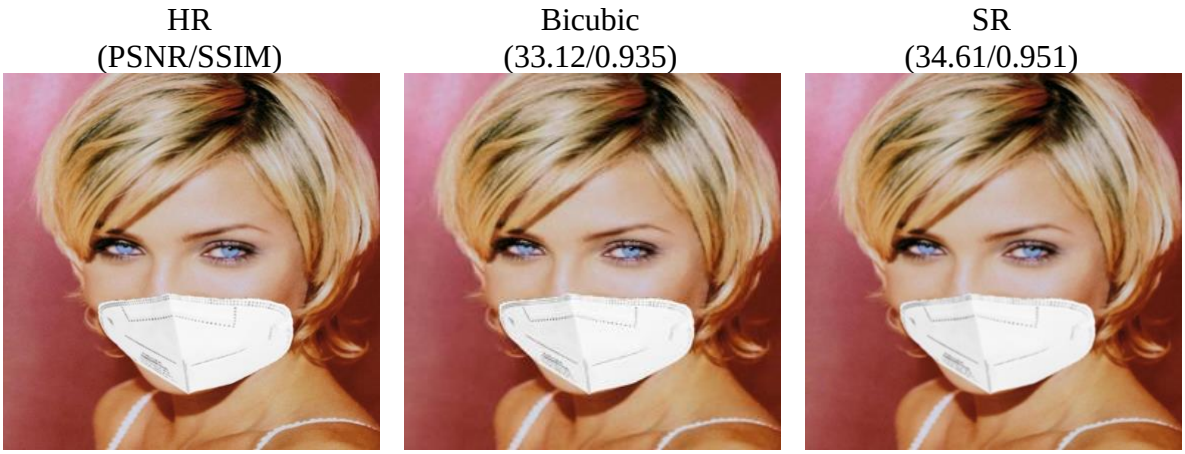


Figure 4-1: FSRCNN 2x upscaling factor against its HR and Bicubic counterparts

Figure 4-2: Comparison of the image inpainting generated output from the 2x SR image against the generated output of its bicubic counterparts and the original image

Figure 4-3: FSRCNN 4x upscaling factor against its HR and Bicubic Counterparts

Figure 4-4: Comparison of the image inpainting generated output from the 4x SR image against the generated output of its bicubic counterparts and the original image

4.1.3. Analysis

As can be seen in Figure 4-1 and Figure 4-3, the SR image had achieved slightly higher PSNR and SSIM results when compared to their bicubic counterparts. It can also be observed that the reconstructed SR image, as seen in Figure 4-2 and Figure 4-4, had shown a slightly better visual and image quality as they are slightly sharper than their bicubic counterparts. However, for the 4x upscaling factor result in Figure 4-3, we could see a prominent checkerboard pattern of artifacts on the mask.

The reason for checkerboard artifacts, as explained by Odena et al. [18], is due to the uneven overlaps during the deconvolutional process. This happens when the output window size, is not divisible by the stride, which is the spacing between the points on the top. While it is possible to reduce artifacts with multiple layers of deconvolutional layers, it is difficult to avoid artifacts from leaking through completely, and they often instead create artifacts of different scales. This is illustrated in Figure 4-5.

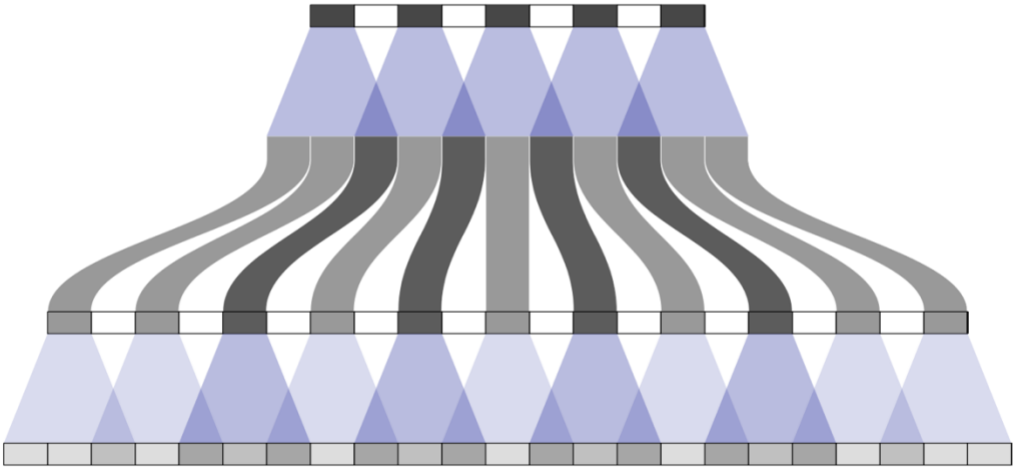


Figure 4-5: Illustration of deconvolutional layer where stide=2 and size=3 (obtained from [18])

Nonetheless, despite the checkerboard artifacts on the SR image of 4x upscaling factor in Figure 4-3, the improvement in PSNR and SSIM on the final generated output by the image inpainting is larger for 4x upscaling than 2x upscaling as can be seen in Figure 4-4. Nevertheless, scrutinizing these experimentation results further, it appears that having a higher PSNR or SSIM does give us a better generated composite image. For instance, the generated output for the bicubic counterpart in Figure 4-4 had left a residue of the mask on the face, thus, giving us obvious visual artifacts and edge trails in the inpainting regions. This contrasts with the generated output for the SR image in Figure 4-2 which gave us a more visually pleasing result without any noticeable colour inconsistency.

4.2. Comparison of Baseline Model with Assembled Model 2

4.2.1. Experiment

In this experiment, we had tested with one of the GAN-based super-resolution techniques, ESRGAN, on the test data to see its effect on the generation of the composite image. In theory, GAN-based SR techniques generally perform better than CNN-based SR techniques as CNN-based SR techniques often output over-smoothed images. As raised by the SRGAN paper [7], the reason for this is due to the usage of MSE loss and PSNR metric that disagrees with capturing visually perceptive attributes, such as having high texture details of an image.

Thus, researchers proposed a new perceptual loss function for GAN, which was described in section 3.2.2.

As depicted in Figure 4-6, there is a noticeable difference between CNN-based SR techniques that uses MSE-based solution, and GAN-based SR techniques that use perceptual loss. The MSE-based solution finds the pixel-wise average of possible solutions, whereas GAN-based solution would drive the image reconstruction towards the natural image [7]. For example, as seen in the following figure, if there are multiple HR images that are a possible solution to the LR image, the MSE loss function would take the average of all the possible HR patches.

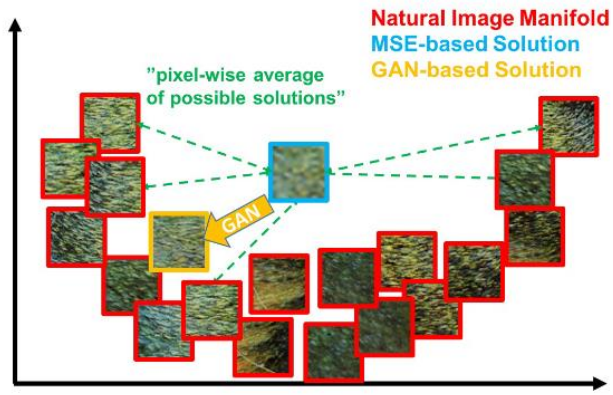


Figure 4-6: Illustration of patches from natural image (red) and SR patches from MSE (blue) and GAN (orange) (obtained from [7])

Hence, in this experiment, we are expecting there to be an improvement in both our quantitative and qualitative evaluation of our output as we have used an improved SR solution. The architecture of assembled model 2 is elaborated in section 3.4.3.

4.2.2. Results

Figure 4-7: ESRGAN 2x upscaling factor against its HR and Bicubic counterparts

Figure 4-8: Comparison of the image inpainting generated output from the 2x SR image against the generated output of its bicubic counterparts and the original image

Figure 4-9: ESRGAN 4x upscaling factor against its HR and Bicubic Counterparts

Figure 4-10: Comparison of the image inpainting generated output from the 4x SR image against the generated output of its bicubic counterparts and the original image

4.2.3. Analysis

As can be seen in Figure 4-7 and Figure 4-9, the SR image had achieved a much larger improvement in the PSNR and SSIM results when compared to their bicubic counterparts. It can also be observed that the reconstructed SR image, as seen in Figure 4-8 and Figure 4-10, had produced better visual and image quality as they are much sharper and has much more natural details than their bicubic counterparts.

Compared to the previous assembled model 1, there is a much larger improvement in the PSNR and SSIM values, and it has a higher textured detail. This may be attributed to the improved GAN-based loss function mentioned previously and as illustrated in Figure 4-6, wherein the GAN-based loss approach drives reconstruction towards the natural image, contrary to the pixel-wise average solution used in CNN-based approaches. On top of that, this model uses RRDB without batch normalization which reduces artifacts significantly.

Similarly as previous, this experimentation has also proven that with a higher PSNR or SSIM, we will get a better generated composite image. The generated output from our SR image for both upscaling factors did not leave any residue of the mask on the face and had given us a visually pleasing result without any noticeable colour inconsistency.

4.3. Comparison of Baseline Model with Assembled Model 3

4.3.1. Experiment

In the previous two experiments, both of the assembled models had achieved better PSNR and SSIM results when compared to the baseline model. However, the generated image under the mask is not sharp.

To fix this, in this experiment, we had extended our model to GFPGAN, which is a super-resolution model that acts as a refinement network for the face. This experiment intends to restore the image into high-quality faces and remove any degradation that may have been caused by the image inpainting network. We hope to achieve a plausible and realistic result of the final output with accurate facial details.

Hence, in this experiment, we are expecting there to be an improvement in the perceptual quality of our output. The architecture of assembled model 3 is elaborated in section 3.4.4.

4.3.2. Results

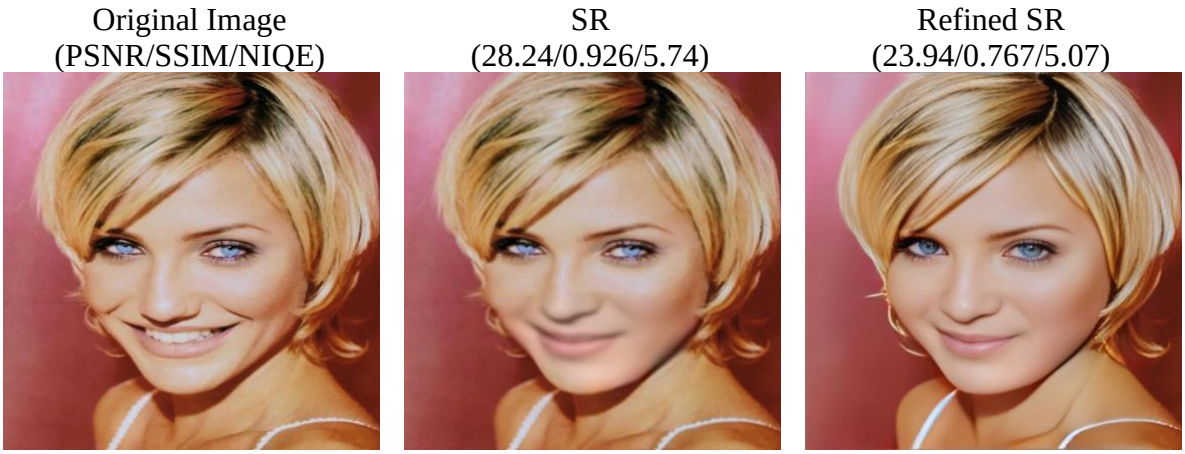


Figure 4-11: Comparison of the image inpainting generated output for ESRGAN 2x and the refined output using GFPGAN

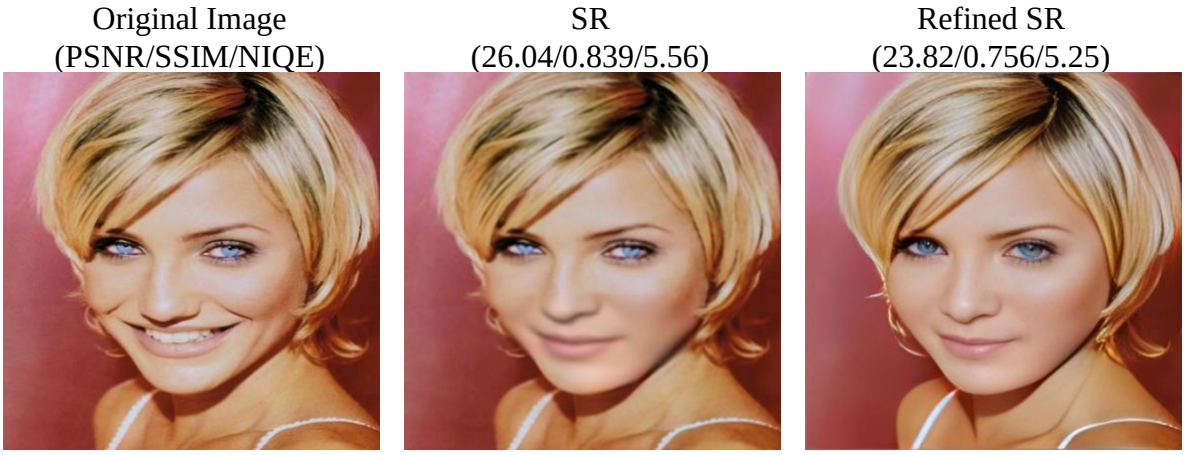


Figure 4-12: Comparison of the image inpainting generated output for ESRGAN 4x and the refined output using GFPGAN

4.3.3. Analysis

As can be seen from the refined images in Figure 4-11 and Figure 4-12 above, the images have managed to successfully meet the objective of the model, which was to remove the degradation on the image and give us an output that is much more plausible and natural. However, it seems that the network had significantly reduced the PSNR and SSIM values as it seems that the network had also altered the colours for enhancement. This is particularly

obvious for the colour of the eyes. A possible reason for this may be because the model does not see such an eye colour as natural, which had caused it to see it as degradation that requires enhancement.

In addition, pixel-wise metrics PSNR and SSIM are not well correlated to the subjective evaluation of human observers [19], hence, these two metrics may not be suitable as an evaluation metric for this experiment. Thus, we have included the NIQE blind evaluation metric to evaluate the quality of the photo. The NIQE evaluation metric has been further elaborated in section 3.1.3. In a nutshell, having a lower value implies a better-quality photo.

As can be seen in Figure 4-11 and Figure 4-12, the refined SR image produced in this experiment has achieved a lower NIQE value than its original SR counterpart. This means that the refined SR image constructed is better in terms of visual quality. We can also observe that the decrease in the NIQE value for the 2x upscaling is much more significant than the 4x upscaling. A reason for this may be due to how there are more pixels and information to be deduced at a higher scaling factor.

Nonetheless, GFPGAN did seem to restore faithful details in the eyes, including in the pupils and eyelashes, as well as the mouth and hair. According to the researchers of GFPGAN, GFP facial prior considers the whole face, rather than separate sections [9]. Thus, we can see that the eyes shape of the generated image may have changed as well if we compared it to the original ground truth image. Our assumption is due to how our model did not generate a mouth with a Duchenne smile, which is a smile that reaches your eyes. Thus, the GFP model had considered the gesture of the mouth of our generated image and consequently restored the eyes to match it.

4.4. Comparison of All Models

4.4.1. Experiment

In this experiment, we are comparing the baseline model, the assembled model 1, the assembled model 2, and the assembled model 3 PSNR, SSIM, and NIQE results on several composite images from the test datasets to evaluate their performance as well as gain an understanding of their potential. In Table 4-4, we have obtained the results by calculating the average of the PSNR, SSIM, and NIQE results on the test datasets.

4.4.2. Results

	Bicubic		AM 1		AM 2		AM 3	
	2x	**4x**	**2x**	**4x**	**2x**	**4x**	**2x**	**4x**
PSNR (dB)	27.37	25.64	27.23	26.32	27.50	27.29	23.70	24.65
SSIM	0.890	0.804	0.883	0.827	0.896	0.853	0.704	0.731
NIQE	5.07	6.20	5.56	5.66	5.27	5.61	5.03	5.00

Table 4-1: Comparison of all models using the evaluation metrics

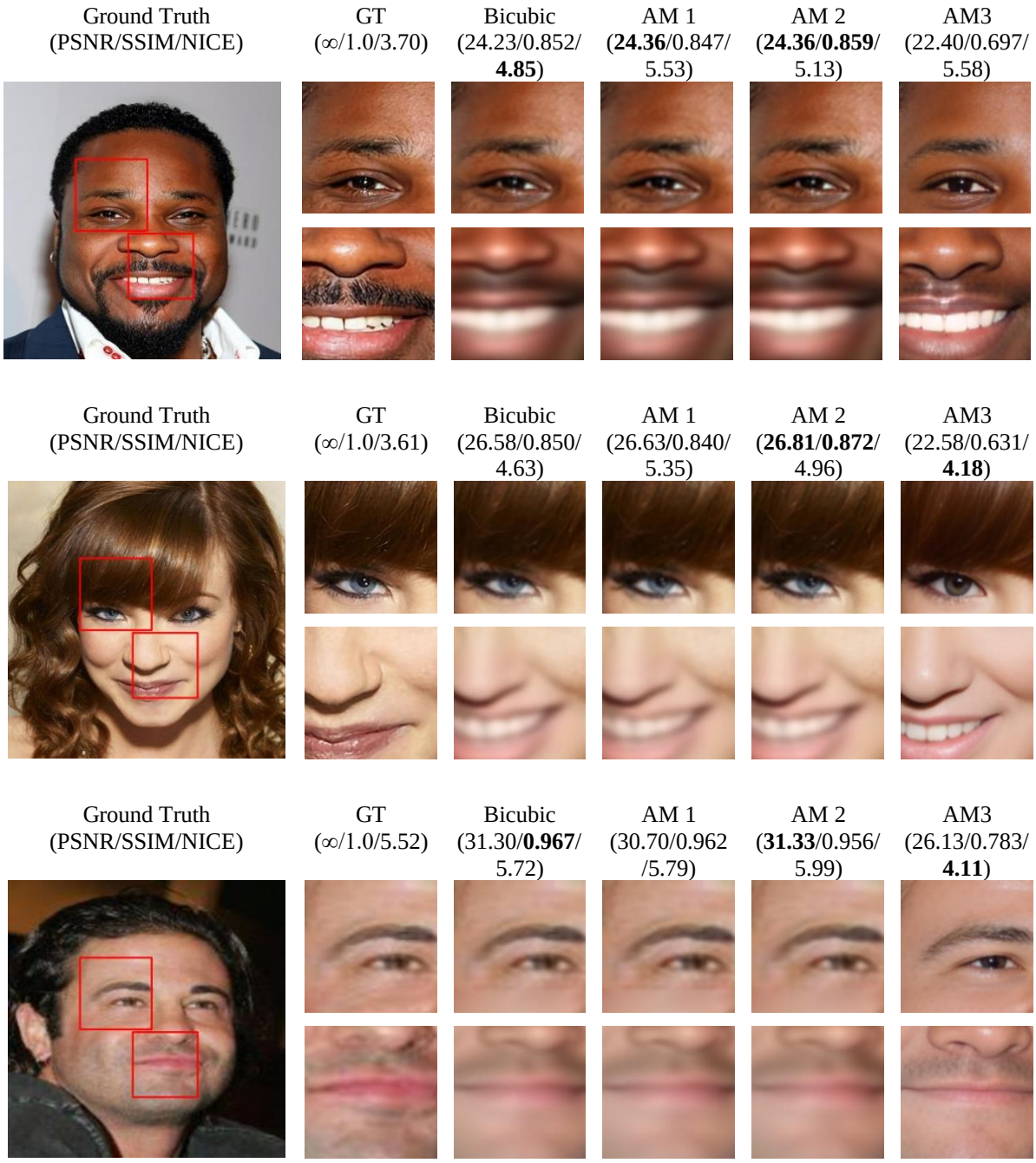


Figure 4-13: 2x upscaling factor performance comparison

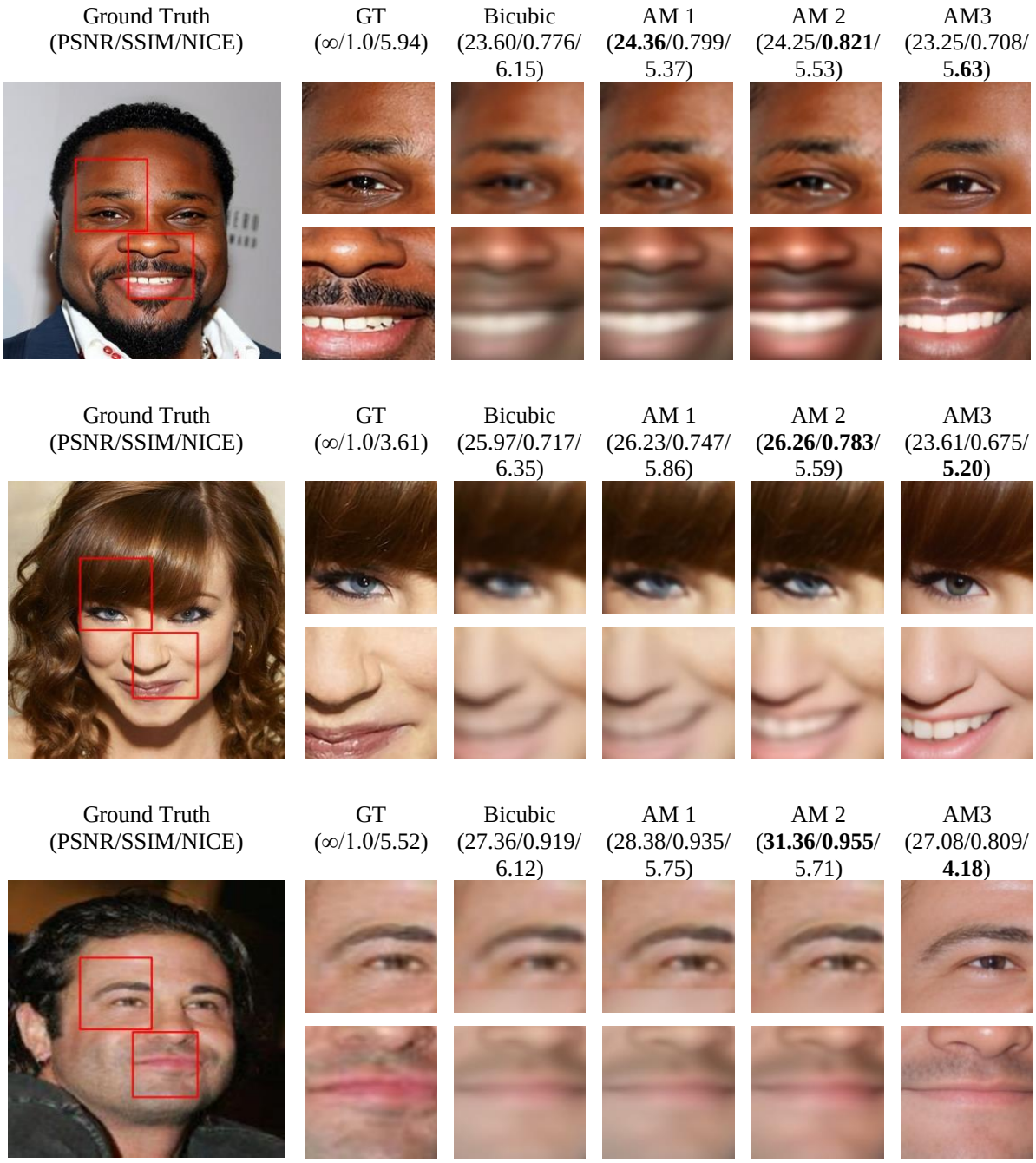


Figure 4-14: 4x upscaling factor performance comparison

4.4.3. Analysis

Figure 4-13 and Figure 4-14 above presents the comparison between all 3 assembled model and the baseline model. It can be observed from Table 4-4 that our assembled model 2 had achieved the best result for PSNR and SSIM compared to the others for both 2x and 4x upscaling factors. On contrary, although assembled model 3 had received the lowest PSNR and SSIM values, they had outperformed the other models with the best NIQE results. From the same table, we could also notice that the model had performed better for 2x upscaling than 4x upscaling. As mentioned previously, a possible reason for this may be due to how there are more pixels and information to be deduced at a higher scaling factor, thus making it challenging for higher upscaling factors to outperform a low upscaling factor.

Based on observation, assembled model 1 had obvious visual artifacts and edge responses from the mask. Meanwhile, assembled model 2 produces better results with no artifacts and is sharper with more natural detail, as seen on the eyelashes and teeth of the second sample image. Hence, it is not surprising that assembled model 2 had achieved better PSNR and SSIM results than assembled model 1.

Although assembled model 3 had attained a considerable amount of enhancement in terms of perceptual quality, it seems to have altered the pixel colours from the original to create a more accurate facial detail, thus causing a significant reduction in the PSNR and SSIM values. Regrettably, it seems that our assembled model 3 had removed the wrinkle details on the first image sample. Nonetheless, regarding achieving our initial project objective, assembled model 3 can generate more detailed composite images, while other models fail to produce enough details (AM 2), or add undesired artifacts (AM 1). Thus, this makes assembled model 3 competitive enough to be used with the objective of our experiment from other models despite the low PSNR and SSIM results.

5. Future Work

Our experiments have only dealt with lower resolution images that were bicubically downsampled from a high-resolution image. This makes our experiment to be only tailored to images that were bicubic downsampled, thus making it limited in terms of scalability and practicability. However, it is not guaranteed that real-world problems use the same downsampling operation. This will inevitably bring about poor performance when the input lower resolution image does not follow what our model was trained for. For example, although our model had given exceptional performance, it may not perform as well on other degradation, such as Gaussian blur or other motion blur and noise kernels. Thus, as future work, we plan to extend our model to be more scalable in terms of handling variations of degradation so that it will be more practicable with real-world data.

On top of that, assembled model 3 seems to give us high accuracy in terms of perceptual quality (as measured by NIQE), but lower in terms of reconstruction accuracy (as measured by PSNR and SSIM). Based on research, it seems that models that perform exceptionally well at minimizing reconstruction error are more inclined to produce unpleasant visual quality. Likewise, models that produce higher-grade visual quality tend to have higher reconstruction errors [20]. Thus, as future work, we plan to evaluate the images on an algorithm that combines both accuracy and perceptual quality for a fair comparison.

6. Conclusion

In this paper, we have explored the use of super-resolution techniques on low-quality images to test if it helps to generate a composite image with greater precision. We proposed a few assembled models that were designed to achieve this objective and had conducted thorough experiments on 2x and 4x upscaling factors.

The first experiment applied FSRCNN as its super-resolution technique and has given us slightly higher PSNR and SSIM results than its bicubic counterpart. Although the image produced is slightly sharper, they have also caused checkboard artifacts to appear on the images. On top of that, the improvement in the image quality does not seem to be very significant which consequently caused there to be artifacts and edge trails responses at the inpainted regions.

The second experiment applied ESRGAN as its super-resolution technique and has given us a much larger improvement in the PSNR and SSIM results than its bicubic counterpart. The improved loss function used to train this model, as well as the usage of RRDB without batch normalization, has caused the image output to contain more comprehensive textured detail and reduced artifacts. Out of all three experiments, this experiment had received the highest PSNR and SSIM value for both 2x and 4x upscaling.

In our third experiment, considering the excellent results from our second experiment, we had improved AM2 by introducing GFPGAN in the final stage for facial resolution enhancement before producing our final output. This experiment had generated the best composite image in terms of visual quality and gave us the best NIQE results. The generated image did not have any artifacts, and they produced an image that is much more natural and believable than the previous two experiments.

From the results we obtained in Section 4, we have concluded that our model does generate a more precise composite image as the quality of the input image improves gradually with the use of super-resolution techniques.

7. Bibliography

[1] S. Farsiu, D. Robinson, M. Elad and P. Milanfar, "Advances and Challenges in Super-Resolution," *International Journal of Imaging Systems and Technology*, vol. 14, 2004.

[2] Z. Wang, J. Chen and S. Hoi, "Deep Learning for Image Super-Resolution: A Survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3365-3387, 2021. doi: 10.1109/tpami.2020.2982166

[3] J. Kamenicky et al., "PIZZARO: Forensic analysis and restoration of image and video data", *Forensic Science International*, vol. 264, pp. 153-166, 2016. doi: 10.1016/j.forsciint.2016.04.027

[4] J. Boyle, "Answer Man: Surveillance video — why is it so bad?", *Citizen-times.com*, 2021. [Online]. Available: https://www.citizen-times.com/story/news/local/2021/03/30/answer-man-why-surveillance-video-so-bad/7044181002/.

[5] C. Dong, C. Loy, K. He and X. Tang, "Image Super-Resolution Using Deep Convolutional Networks", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, 2016. doi: 10.1109/tpami.2015.2439281.

[6] C. Dong, C. C. Loy, and X. Tang, “Accelerating the super-resolution convolutional neural network,” *Computer Vision and Pattern Recognition*, 2016. doi: 10.48550/ARXIV.1608.00367.

[7] C. Ledig et al., “Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network.” *Computer Vision and Pattern Recognitio*n, 2017. doi: 10.48550/ARXIV.1609.04802.

[8] X. Wang et al., “ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks.” *Computer Vision and Pattern Recognition*, 2018. doi: 10.48550/ARXIV.1809.00219.

[9] X. Wang, Y. Li, H. Zhang, and Y. Shan, “Towards Real-World Blind Face Restoration with Generative Facial Prior.” *Computer Vision and Pattern Recognition*, 2021. doi: 10.48550/ARXIV.2101.04061.

[10] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and Improving the Image Quality of StyleGAN.” *Computer Vision and Pattern Recognition*, 2019. doi: 10.48550/ARXIV.1912.04958.

[11] “NET: Convolutional Networks for Biomedical Image Segmentation,” *Uni Freiburg*. [Online]. Available: https://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/.

[12] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, “Image Quality Assessment: From Error Visibility to Structural Similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, April 2004.

[13] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a ‘completely blind’ Image quality analyzer,” *IEEE Signal Processing Letters*, vol. 20, no. 3, 2013. doi: 10.1109/lsp.2012.2227726

[14] Zhang Y, Yu W, “Comparison of DEM Super-Resolution Methods Based on Interpolation and Neural Networks” *Sensors*, vol. 22, no. 3, 2022. doi: https://doi.org/10.3390/s22030745

[15] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled Faces in the Wild: A Database for Studying,” *University of Massachusetts, Amherst*, Oct-2007. [Online]. Available: http://vis-www.cs.umass.edu/lfw/.

[16] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive Growing of GANs for Improved Quality, Stability, and Variation.” *Neural and Evolutionary Computing*, 2017. doi: 10.48550/ARXIV.1710.10196.

[17] A. Anwar and A. Raychowdhury, “Masked Face Recognition for Secure Authentication,” *ArXiv*, vol. abs/2008.11104, 2020.

[18] A. Odena, V. Dumoulin and C. Olah, “Deconvolution and Checkerboard Artifacts,” *Distill,* 2016.

[19] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, “The 2018 PIRM Challenge on Perceptual Image Super-resolution,” *Computer Vision and Pattern Recognition,* 2018, doi: 10.48550/ARXIV.1809.07517.

[20] Y. Blau and T. Michaeli, “The Perception-Distortion Tradeoff,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE*, Jun. 2018. doi: 10.1109/cvpr.2018.00652.