The University of Hong Kong
COMP 4801 Final Year Project
Detailed Interim Report
January 20, 2019
Topic: Face Image Super-Resolution using Deep Learning

Supervisor: Dr. Kenneth K.Y. Wong
Team Name: DeepStrive
Team Members: Lo Tsz Fung (3035270953)
Tsang Suet Ying Florence (3035272341)
Wong Man Yu (3035274313)
Abstract

Super-resolution is the reconstruction of high-resolution images from low-resolution images. Face image super-resolution is useful for security, surveillance and multimedia fields. The latest advancement in deep learning technology opens possibilities for more plausible face image super-resolution by learning from deeper facial features. This project aims at experimenting and testing various deep learning algorithms for face image super-resolution, as well as improving existing algorithms. Replication of several state-of-the-art super-resolution models has been completed and baseline models have been created. Several face specific super-resolution models have been proposed, namely the LandmarkNet, the FeaturemapNet, the heatmapNet and the SRFaceNet. These models incorporated various face-related information such as facial landmark and recognition loss. Further work in this project is needed on fine tuning hyperparameters, network architecture and loss function. If time allows, the team would implement other tasks related to super-resolution.
Acknowledgements

We would like to take this opportunity to thank our supervisor Dr. Kenneth K.Y. Wong from the Department of Computer Science for his advice and careful guidance in this project, as well as the permission to use the GPU from his research group.
Contents

Abstract ........................................................................................................................................................................ i
Acknowledgements ........................................................................................................................................................... i
List of Figures ............................................................................................................................................................... iv
List of Tables ................................................................................................................................................................. iv
List of Abbreviations ...................................................................................................................................................... v
1. Introduction ................................................................................................................................................................. 1
   1.1. Background and Motivation ........................................................................................................................................ 1
   1.2. Objectives ................................................................................................................................................................. 2
   1.3. Scope ....................................................................................................................................................................... 3
   1.4. Outline of Report ...................................................................................................................................................... 3
2. Related Works ............................................................................................................................................................... 4
3. Methodology ................................................................................................................................................................. 6
   3.1. Datasets .................................................................................................................................................................. 6
      3.1.1. CelebA ........................................................................................................................................................... 6
      3.1.2. Helen Dataset .............................................................................................................................................. 6
   3.2. Development methods .............................................................................................................................................. 6
4. Discussion of results ....................................................................................................................................................... 8
   4.1. General super-resolution models .................................................................................................................................. 8
      4.1.1. Replication of previous models ......................................................................................................................... 8
      4.1.2. Design Considerations ..................................................................................................................................... 8
      4.1.3. Benchmark models ......................................................................................................................................... 9
      4.1.4. Results .......................................................................................................................................................... 10
   4.2. Face image super-resolution models .......................................................................................................................... 11
      4.2.1. Incorporation of facial landmarks ................................................................................................................... 11
      4.2.2. Incorporation of face recognition .................................................................................................................. 15
      4.2.3. Evaluation ................................................................................................................................................... 17
5. Challenges ................................................................................................................................................................... 18
   5.1. Downsampling method dependency .......................................................................................................................... 18
   5.2. Difficulty in achieving state-of-the-art performance ............................................................................................ 18
5.3. Difficulties of model evaluation ........................................................................ 19

6. Future Planning .................................................................................................... 20
   6.1. Future development ....................................................................................... 20
   6.2. Optional features implementation .................................................................. 20
   6.3. Schedule ........................................................................................................ 20

7. Conclusion ........................................................................................................... 22

References ............................................................................................................... 23
List of Figures

Figure 1: Description of the development methods ................................................................. 7
Figure 2: An illustration of our two benchmark residual networks .......................................... 9
Figure 3: Output of general super-resolution models. From left to right: low-resolution image, bicubic interpolated image, SRCNN, SRGAN, ResNet1, ResNet2, and the ground truth .......... 10
Figure 4: Illustration of the FeaturemapNet incorporating feature map into the model ............. 11
Figure 5: Examples of feature maps generated by different methods ........................................ 12
Figure 6: First attempt to incorporate landmark information by multi-task learning ............... 12
Figure 7: Network structure of the proposed LandmarkNet ...................................................... 13
Figure 8 Qualitative Comparisons of LandmarkNet with benchmark algorithm .................. 14
Figure 9: An illustration of SRHeatmapNet ........................................................................... 14
Figure 10: An illustration of the SRFaceNet model ............................................................... 15
Figure 11: Qualitative Comparisons of various SRFaceNets with the benchmark algorithm .... 16
Figure 12: Failure to produce good result when downsampling method of LR image is different from training data ................................................................. 18

List of Tables

Table 1: PSNR metrics for general super-resolution models.................................................... 10
Table 2 Quantitative comparisons on LandmarkNet with benchmark algorithm .................. 14
Table 3: Quantitative comparisons on PSNR among SRFaceNet and the benchmark algorithm 16
Table 4: Qualitative comparisons on top-n face recognition accuracy among SRFaceNet and the benchmark algorithm .............................................................................. 17
Table 5: Quantitative comparisons on face image super resolution models with benchmark algorithm ........................................................................................................... 17
Table 6: Project schedule ......................................................................................................... 21
List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>Convolutional Neural Network</td>
</tr>
<tr>
<td>GAN</td>
<td>Generative Adversarial Nets</td>
</tr>
<tr>
<td>GPU</td>
<td>Graphical Processing Unit</td>
</tr>
<tr>
<td>ResNet</td>
<td>Residual Network</td>
</tr>
<tr>
<td>SR</td>
<td>Super-resolution</td>
</tr>
<tr>
<td>SISR</td>
<td>Single-Image Super-resolution</td>
</tr>
<tr>
<td>MISR</td>
<td>Multiple-Image Super-resolution</td>
</tr>
<tr>
<td>LR</td>
<td>Low-resolution</td>
</tr>
<tr>
<td>HR</td>
<td>High-resolution</td>
</tr>
<tr>
<td>MSE</td>
<td>Minimum Squared Error</td>
</tr>
<tr>
<td>PNSR</td>
<td>Peak Noise to Signal Ratio</td>
</tr>
</tbody>
</table>
1. Introduction

1.1. Background and Motivation

High-resolution human face images are important in various fields. Particularly, in security and surveillance fields, high-resolution face images are required for face recognition or detection purpose [1]. The multimedia industry also desires high-resolution face images since they provide better audience perception and satisfaction [2].

In reality, the resolution of face images may not necessarily be as high as desired. These may sometimes be limited by hardware of imaging devices [2], or the fact that the image is sometimes out of focus. Image resolution may also be limited by the distance between the human subject and the imaging device, such that the face image taken is small [2]. Therefore, it is highly desirable and valuable if high-resolution face image could be reconstructed from low-resolution image using super-resolution techniques. The latest advancement in deep learning technology opens possibility for more plausible image super-resolution by learning from realistic facial features.

There are several rooms for improvements on existing solutions of super-resolution by deep learning and algorithmic models. In terms of:

1) Accuracy and perceptual quality of outputs
2) Computational resources required, such as time and memory usage

Moreover, existing solutions for super-resolution are usually more generic, without specializing on a category of images. It is believed that studying face image super-resolution could bring new value to the field of image super-resolution.
1.2. Objectives

This project aims at implementing a deep neural network for super-resolution with a narrower input scope, which is face images. By experimenting with various deep learning algorithms and improving existing algorithms, we expect that our model will produce photo-realistic outputs while reducing computation power and being able to handle multi-scale super-resolution. In addition, the output image is expected to be a high-resolution face image of the same person.
1.3. Scope

Image super-resolution falls into two categories - multiple-image super-resolution (MISR) and single-image super-resolution (SISR). MISR refers to the fusing of information from multiple low-resolution images to produce a high-resolution face image [3]. On the other hand, the input for SISR is one single low-resolution image, from which the resolution is enhanced by previous learning of the relationship between the low-resolution examples of the same category (e.g. face images) and their high-resolution counterparts [4].

Our project will mainly focus on single-image super-resolution for face images. We believe that SISR is a more valuable and challenging problem as the outcome is more dependent on the capability of the algorithm to reconstruct image information missing in the low-resolution single image source. Additionally, data for SISR is more readily available. The training and testing data for SISR are independent face images, while that for MISR are sets of images of the same face, the source of which is scarcer. A rich data source will ease our data collection process and will provide better grounds for training and experimenting on the deep learning model.

Furthermore, our project will mainly focus on frontal face images which provides the most facial features. Limiting the scope to only frontal face images should also make the deep learning model more specialized and mature in the task.

1.4. Outline of Report

The remainder of the paper is organized as follows. Chapter 2 provides a literature review on the recent super-resolution algorithms and some evaluation methods for the algorithms. Chapter 3 gives a detailed description of the methodology used in the project. It describes our datasets and how we use them in our project. Several milestones for our model development have also been established in this chapter. Chapter 4 presents the current status of the project. First, we report our work done on the development of both general and face-specific super-resolution models. Several experiments and design considerations are highlighted. After that, we make both qualitative and quantitative comparison among our models and the benchmark models. Chapter 5 discusses various difficulties we faced in the project and suggests several possible solutions to our challenges. Chapter 6 conclude the report by restating our major progress and findings at the current stage as well as discussing the future steps for the project. Finally, our project schedule is presented.
2. Related Works

Traditional approaches of super-resolution generate high-resolution (HR) images from low-resolution (LR) images without the use of deep learning. Interpolation is one of the basic approaches in which the interpolated pixels are calculated from neighbouring LR pixel values [5]. Several other SISR algorithms focused on the reconstruction of sharp edges in HR images, or the utilization of image statistics or exemplar image patches for generating HR images [5]. However, in these models, high frequency details of the images are lost.

Recently, deep learning techniques have been used extensively in SISR. This enables the reconstruction of more realistic results and finer details. Dong et al. [6] introduced a three-layer SRCNN, which extracts features from the input image, performs a non-linear mapping, and reconstructs an HR image.

Some studies aimed at improving the model efficiency [7, 8] while some others focused on the model accuracy. Kim et al. [9] made use of a recursive layer to make the network deeper such that it is capable to enlarge the receptive field, which provides more contextual information for learning. In Kim et al. [10], a very deep network that comprises 20 layers brought about notable gain in the accuracy. Although deeper networks can be more powerful, they are more difficult to be trained due to the vanishing gradient problem [11]. Residual learning, gradient clipping and skip connections are some popular techniques to mitigate this issue [10, 12, 13, 14, 15, 16].

Among these models, Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR) are the common choices of the loss function and evaluation metrics respectively. A common issue is that although minimizing MSE tends to give a higher PSNR, it does not necessarily reflect the perceptual quality of the output HR image since MSE often overly smoothens the fine details in the images [17].

Ledig et al. [17] proposed a generative adversarial network (GAN) and a perceptual loss that combines adversarial loss and content loss. Despite that the PSNR stays the same, they demonstrated finer HR output images.

More recently, to deal with the instability of GANs, Dahl et al. [18] extended the pixelCNNs and developed a new probabilistic model for SISR, which consists of a prior network and a conditioning network. Using a log-likelihood function for training, the prior network can potentially learn the prior information of a particular type of images such as face images. Additionally, a study involving human observers was conducted to evaluate the performance of the model in generating realistic images.
On the other hand, the model would be more useful if it is able to support different scaling factors from the input image to the output image. Some multi-scale models achieve this by enhancing the image resolution along the network and at the same time producing intermediate outputs of various resolutions [10, 13, 15].

It is also worth noting that some models use bicubic-upsampled images as input and train the mapping function between the inputs and HR images [6, 9, 10], while other models take LR images as input and directly learn the upscaling filters [8, 15, 19] inside the network.

All the above related works provide insights for us on formulating our project strategy. It will also guide us in exploring which technologies or approaches could be re-used or improved.
3. **Methodology**

The project is divided into four stages – 1) image collection and pre-processing; 2) results replication and analysis; 3) model creation, evaluation and improvement cycle, and 4) optional features implementation. The sections below describe some of the important stages in details.

3.1. **Datasets**

The team mainly utilized two face image datasets, namely CelebA and Helen Dataset for various process including training, evaluation and testing multiple features.

3.1.1. CelebA

CelebA is a facial image dataset containing 202,599 facial images. All face images are in near-frontal poses, cropped to the same aspect ratio, and aligned. The aligned and cropped data could ease the training of our model as they allow the model to easier recognize the location of different facial features. The dataset has been divided into 3 partitions, 182,599 images for training, 10,000 for evaluation and the remaining 10,000 for testing.

Apart from the face image data, this dataset also contains 5 facial landmarks (eye, nose and mouth) of every face. In this project, we mainly use this dataset for the training and evaluation process.

3.1.2. Helen Dataset

Helen Dataset is a facial image dataset containing 2,330 face images with 11 landmark heatmaps. The main task to be achieved with this dataset is to train a landmark heatmap prediction network and incorporate that into deep learning models.

3.2. **Development methods**

The diagram below describes the development methods. The team intend to build new models by incorporating the approaches in various existing models and adding some new ideas based on our insights. Possible directions include incorporating face images prior into our model, i.e. allowing the model to learn the prior information of the specific category of images - face image, employing appropriate neural network architectures from previous studies, or modifying the objective function to be trained.

After creating a deep learning model, the team trains the model on about 180,000 celebA face images for a maximum of 30 epoches. Then the model is evaluated both qualitatively and
quantitatively using metrics like MSE and PSNR. The performance of our models with the models from research papers are then compared.

The development is an iterative process. Through all the evaluations, insights would be gained on how to further improve our model. Another version of the model is created and the cycle is repeated until a reasonably satisfactory model is created.

Figure 1: Description of the development methods
4. Discussion of results

4.1. General super-resolution models

This section summarizes our work done on the development of super resolution models for general images.

4.1.1. Replication of previous models

Before starting the development of our own general models, the team first replicated models that have been developed previously by other researchers. In particular, using the CelebA face image dataset, the results of SRCNN [6] in 2015 as well as SRResNet and SRGAN [17] in 2017 are reproduced. We built a pipeline for data input, model development and training. Moreover, feature maps of some intermediate layers are visualized for analysis. After this stage, we have already gained a basic understanding on how deep learning models could be used for super-resolution.

4.1.2. Design Considerations

With reference to several state-of-the-art models such as SRResNet, we decided to use residual networks, which are formed by stacks of residual blocks, as our backbone network. With this backbone network, experiments are conducted on various aspects of deep learning models in order to gain more understanding towards their impacts on the SR performance.

Normalization methods

While we can directly input the RGB image to the model, several normalization methods such as scaling to [0,1] or [-1,1] are common in image processing. Image standardization that subtracts the mean from the image and then divides the result by the standard deviation of the image is another way of image normalization. In our project, all inputs are scaled to [-1,1] before passing to our neural networks. The advantage of this method is that it centres the data at zero and within a small range, allowing the training to converge faster.

Upscaling methods

In addition, various ways of upsampling low-resolution images have been considered. For instance, our model could either take an image that has been upsampled by bicubic interpolation as input or specifically upsample the image inside using transposed convolution or subpixel convolution.
Residual Blocks

Moreover, several different residual blocks have been tested for building our residual network. Particularly, we implemented the basic residual block and its bottleneck version introduced in [23], as well as the residual block that uses identity mappings [24].

Network Architecture

For the network architecture, the number of residual blocks and the position of upscaling filters in our network were varied. Different parameters such as filter sizes and output channels were also experimented. As our model is expected to be trained on one GPU, we should control the parameters and the number of residual blocks such that our model can fit in the memory of one GPU. On the other hand, it was found that residual blocks after upscaling filters were needed to produce clear outputs.

Other considerations include the activation function, the optimizer and other hyperparameters such as learning rate. In general, the team will follow the convention in our project, for example, using Rectified Linear Unit (ReLU) activation function and Adam Optimizer.

4.1.3. Benchmark models

Figure 2: An illustration of our two benchmark residual networks. ResNet1 on the left requires input to be a bicubic upsampled image while ResNet2 on the right upsamples images inside the neural network.
ResNet1

The illustration of ResNet1 is shown on the left of Figure 2. It takes a bicubic upsampled LR image as input. After applying a convolutional layer, 10 residual blocks with different kernel sizes are used. Finally, a convolutional layer that has a tanh activation acts as the output layer.

ResNet2

On the other hand, ResNet2 which is on the right side of Figure 2 upsamples the original 22x27 LR image inside the network. The input image is first passed through a convolutional layer and three residual blocks for pre-processing, and then two sets of transposed convolutional layers with residual blocks are used to produce output images of 4 times larger resolution. Moreover, we add a skip connection for every three residual blocks.

4.1.4. Results

Figure 3: Output of general super-resolution models. From left to right: low-resolution image, bicubic interpolated image, SRCNN, SRGAN, ResNet1, ResNet2, and the ground truth.

Table 1: PSNR metrics for general super-resolution models

<table>
<thead>
<tr>
<th>Model</th>
<th>Bicubic</th>
<th>SRCNN</th>
<th>SRGAN</th>
<th>ResNet1</th>
<th>ResNet2</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>23.32</td>
<td>23.64</td>
<td>25.77</td>
<td>24.76</td>
<td>25.13</td>
</tr>
</tbody>
</table>

Figure 3 shows a qualitative comparison of several general super resolution models that have been replicated and created. The first SR deep learning model SRCNN does not produce satisfactory results. It is not capable of recovering the landmarks and generating a clear face. In contrast, SRGAN and both of our models ResNet1 and ResNet2 could reconstruct the face from LR images. While our residual networks that use an MSE loss function generate smooth images, SRGAN seems to be able to produce fine details in the images, leading to outputs that are even
clearer than the truth. Yet, it is questionable whether faces in the images generated by these general SR models remain unchanged.

Table 1 presents a quantitative comparison among the same models. The result is consistent with our observation. SRGAN achieves the highest PSNR and our residual networks are slightly worse while the performance of bicubic interpolation and SRCNN are rather poor. Among our two ResNet, ResNet2 that utilizes upscaling filters to upsample images in the network attains a better PSNR than its counterpart.

4.2. Face image super-resolution models

The above general models could be used in general super-resolution. Although they give plausible results for face-image super-resolution, they fail to incorporate face-related information into the super-resolution process. We observe that when these models are used alone, the position of some facial features are inaccurate and the identity of the face might not be preserved.

To mitigate this, facial information such as landmark, feature map, heatmap and recognition loss were incorporated into our models. Several models incorporating different facial features have been built and evaluated.

4.2.1. Incorporation of facial landmarks

Facial landmarks provide the accurate locations of facial key points. It is believed that incorporation of landmark information into our models allows more accurate recovery of facial features. Note that celebA dataset only has a ground truth of 5 landmarks. In our implementation, face alignment libraries like [27] and dlib was used to estimate the 68 landmarks as the ground truth. Multiple methods are used in order to incorporate the facial landmark information in the baseline model above.

FeaturemapNet

![Figure 4: Illustration of the FeaturemapNet incorporating feature map into the model](image-url)
This model simply utilizes resnet1 as described above as the baseline model. The resnet1 is altered to accept image with 4 channels while other implementation details remain the same, in which the first 3 channels are for the original LR image and the last channels is for the manually created feature maps that is basically a grayscale image with the region around the facial landmark being highlighted. The two images are concatenated in order to be used as the input data.

![Figure 5: Examples of feature maps generated by different methods](image)

With the source of ground truth facial landmarks like 68 landmarks predicted by dlib and the 5 landmarks as provided in celebA dataset, we have experimented with different methods to create the feature maps like the two figures shown above. By telling the model the location of the facial landmarks, it is hoped that the model would achieve better performance in the SR task.

**LandmarkNet**

![Figure 6: First attempt to incorporate landmark information by multi-task learning](image)

It was attempted to perform multi-task learning by predicting landmark as an intermediate output. It was believed that this network could restrict the first few layers to learn the high level and face-related features, instead of only learning to do basic interpolation and upsampling.

However, after inspecting the performance of the above model (PSNR = 24.68), it does not show improvement over the baseline model (PSNR = 24.76). It was believed that the LR image information was lost after passing through the three convolution layers. The problem was resolved by feeding the LR image back to the network at the middle layer. This improved model is the LandmarkNet.
The LandmarkNet consists of two parts: *SR network* and *landmark detection network*. The first three layers of the *landmark detection network* is same as the three convolutional layers of SRCNN described in [6]. Then a flatten layer followed by a fully connected layer is used to estimate the landmark information as a 136-dimensional vector which represents the coordinates of 68 facial landmarks. The *SR network* has the same architecture as the ResNet1 described in section 4.1.3.

Denote $x$ as the low-resolution input image, $y$ and $k$ as the super-resolution image and the estimate landmark information. Define prior $p$ as the activation of the *landmark detection network* at the layer just before the last FC layer.

Let $P$ denote the mapping from an LR image $x$ to a prior $p$. And landmark $k$ could be estimated from $p$ through a fully connected layer $K$. Then,

$$
p = P(x)$$

$$k = M(x) = K(p) = K(P(x))$$

Where $M$ denotes the mapping from LR image $x$ to landmark $k$. After obtaining $p$, the *SR network* $F$ is utilized to recover the SR image by concatenating the LR image $x$ and the prior $p$.

$$y = F(x, p)$$

Given a training set of $N$ examples $\{x^{(i)}, \tilde{y}^{(i)}, \tilde{k}^{(i)}\}_{i=1}^N$, where $\tilde{y}^{(i)}$ is the ground-truth HR image of the LR image $x^{(i)}$ and $\tilde{k}^{(i)}$ is the corresponding ground truth landmark information, LandmarkNet has the loss function,

$$L(\Theta) = \frac{1}{N} \sum_{i=1}^{N} \left( \left\| \tilde{y}^{(i)} - y^{(i)} \right\|^2 + \alpha \left\| \tilde{k}^{(i)} - k^{(i)} \right\|^2 \right)$$

Where $\Theta$ denotes the parameter set, $\alpha$ is the weight of the landmark loss and $y^{(i)}$, $k^{(i)}$ are the recovered HR image and estimated landmark information of the $i$-th image respectively.
The model was trained using Adam Optimizer with learning rate of 0.001 and epsilon of 1e-7. The batch size was set to be 32 and $\alpha$ was set to be 1e-3. Quantitatively, it achieves better PSNR than ResNet1. Qualitatively, the locations of facial features of the face images produced by LandmarkNet are more accurate than that by ResNet1.

![Image](image_url)

Figure 8 Qualitative Comparisons of LandmarkNet with benchmark algorithm. Note that the landmark locations of the face images produced by LandmarkNet are more accurate than that by ResNet1.

Table 2 Quantitative comparisons on LandmarkNet with benchmark algorithm

<table>
<thead>
<tr>
<th></th>
<th>Bicubic</th>
<th>ResNet1</th>
<th>LandmarkNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>23.32</td>
<td>24.76</td>
<td>25.02</td>
</tr>
</tbody>
</table>

Face Heatmap

![Image](image_url)

Figure 9: An illustration of SRHeatmapNet. The model consists of two networks, namely the heatmap regression network and the super-resolution network.
As shown in Figure 9, this model comprises a heatmap regression network and a SR network. The heatmap regression network employs an Hour-Glass structure described in [25], which aims to extract features at every scale. Its main characteristic is the use of a number of max pooling layers to downsample the input and then a number of transposed convolutional layers to upsample the result again. There are also skip layers that directly process the image at different resolution. The results of these skip layers are then combined with the features at the corresponding resolution extracted after downsampling and upsampling. After the original resolution is restored after a series of downsampling and upsampling, two residual blocks and a convolutional layer are used to produce the heatmap outputs. For the SR network, it is the same as our ResNet1 except that the outputs of its second residual blocks are added with the outputs of the last residual block of the heatmap regression network.

The two networks in this model are jointly trained with intermediate supervision, meaning the loss function has two terms. The first one come from the mean square error from the SR network. The second is a per-pixel sigmoid cross-entropy loss from the heatmap regression network as the heatmap can be viewed as a probability map on which every pixel performs a binary classification to see whether it belongs to a certain landmark feature.

4.2.2. Incorporation of face recognition

As one of the objectives of our model is to preserve the identity of the person in the images, an intuitive direction would be the incorporation of face recognition task into our model.

Model details

Figure 10 provides a high-level overview of the model SRFaceNet. The SR network is simply a ResNet1 shown in chapter 4.1.3. Then we feed the resulting SR image and the truth to the pretrained state-of-the-art face recognition network FaceNet [26] to obtain two 512-dimensional feature vectors $v_{SR}$ and $v_{truth}$ for the face in the two images. As the Euclidean distance of two vectors given by FaceNet can be regarded as a measure of face similarity, we include this distance in the loss function. Hence given a training set of N samples, the overall loss function is given by $L = \frac{1}{N} \sum_{i=1}^{N} (||\tilde{y}^{(i)} - y^{(i)}||^2 + \alpha \cdot ||v_{SR}^{(i)} - v_{truth}^{(i)}||)$ where $\tilde{y}$ and $y$ are the truth and generated SR image respectively and $\alpha$ is a hyper-parameter controlling the proportion of the
recognition loss in the total loss function. Then we perform gradient descent and update only weights in the SR network while keeping the FaceNet unchanged. With this loss function, the solution space can probably be more constrained so that faces in the super-resolved images could be closer to the original faces.

**Experiments**

The hyper-parameter $\alpha$ is adjusted so that the recognition loss contributes to the loss function to different extents. Specifically, it has been set to $1e^{-4}$, $2e^{-4}$ and $4e^{-4}$ in SRFaceNet1, SRFaceNet2 and SRFaceNet4 respectively.

**Results**

Perceptual quality and PSNR

![Qualitative Comparisons of various SRFaceNets with the benchmark algorithm.](image)

Figure 11: Qualitative Comparisons of various SRFaceNets with the benchmark algorithm.

<table>
<thead>
<tr>
<th>Model</th>
<th>ResNet1</th>
<th>SRFaceNet1</th>
<th>SRFaceNet2</th>
<th>SRFaceNet4</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>24.76</td>
<td>24.67</td>
<td>24.68</td>
<td>24.55</td>
</tr>
</tbody>
</table>

Table 3: Quantitative comparisons on PSNR among SRFaceNet and the benchmark algorithm.

The results of various SRFaceNets are demonstrated in figure 11. It can be observed that images produced by SRFaceNets are even smoother than the benchmark Resnet1. Among SRFaceNets, the best output is given by SRFaceNet1 which sets $\alpha$ to be $2e^{-4}$. Both SRFaceNet2 and SRFaceNet4 generates more blurry images. Furthermore, the PSNR of SRFaceNet models are also slightly lower than that of ResNet (See Table 3), which substantiates the observation that the output perceptual quality is not as satisfactory as desired.
Face recognition accuracy

Table 4: Qualitative comparisons on top-n face recognition accuracy among SRFaceNet and the benchmark algorithm

<table>
<thead>
<tr>
<th>n</th>
<th>ResNet1</th>
<th>SRFaceNet1</th>
<th>SRFaceNet2</th>
<th>SRFaceNet4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.8704</td>
<td>0.9203</td>
<td>0.9375</td>
<td>0.9368</td>
</tr>
<tr>
<td>5</td>
<td>0.9291</td>
<td>0.9624</td>
<td>0.9707</td>
<td>0.9704</td>
</tr>
<tr>
<td>10</td>
<td>0.9463</td>
<td>0.9712</td>
<td>0.9782</td>
<td>0.9783</td>
</tr>
</tbody>
</table>

Since the purpose of the incorporation of face recognition task is to preserve the identity in the images, it is reasonable to test if the incorporation successfully achieves this goal. Here we use a top-n recognition accuracy as a metrics. Given a SR image, we compute a feature vector for the face, compare its Euclidean distance with all feature vectors from the truth in our database and see whether the original image is among one of the n closest vectors. The recognition accuracy simply finds the percentage that the ground truth appears in one of the top n predictions from the SR images. The experimental results presented in Table 4 show that the incorporation of face recognition task indeed helps preserve the identity. For example, for n=5, without recognition loss, the recognition accuracy is only 92.91%. Setting $\alpha=1e^{-4}$, the accuracy increases significantly to 96.24%. The accuracy is even higher (97.07%) for $\alpha=2e^{-4}$. Yet, an even larger $\alpha$ seems not to be useful as the accuracy drops slightly to 97.04%. Overall, including a recognition loss seems to be helpful to make sure that the face does not change drastically during super-resolution.

4.2.3. Evaluation

This section provides the evaluation metrics and the image output of the model to compare between the models.

It can be observed that by incorporating face prior information into the deep learning model, the performance on PSNR and loss improves, which indicates that the models is able to learn from the landmark information. Also, by enabling end-to-end training of landmark location/heatmap and SR can further improve the performance of the model. Since the model can capture the information given by the facial landmarks.

Table 5: Quantitative comparisons on face image super resolution models with benchmark algorithm

<table>
<thead>
<tr>
<th>Model</th>
<th>PSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>RESNET1</td>
<td>24.76</td>
</tr>
<tr>
<td>FeaturemapNet</td>
<td>24.85</td>
</tr>
<tr>
<td>LandmarkNet</td>
<td>25.02</td>
</tr>
<tr>
<td>SRHeatmapNet</td>
<td>26.22</td>
</tr>
<tr>
<td>SRFaceNet</td>
<td>24.68</td>
</tr>
</tbody>
</table>
5. Challenges

5.1. Downsampling method dependency

To obtain training data for our model, the low-resolution images are artificially constructed from high-resolution images. In this way, the super-resolved low-resolution images could be evaluated by comparing with the original high-resolution image. There are various downsampling methods such as bilinear resampling and bicubic interpolation. It was observed that the model fails to produce good result when downsampling method of low-resolution image is different from training data. For example, SRGAN is trained using bicubic downsampled LR images. If bilinear LR image is used to test the model, it fails to produce good results.

![Figure 12: Failure to produce good result when downsampling method of LR image is different from training data](image)

This implies that the model would very likely fail to produce good result when applied to real-world low-resolution, low quality images. Since the ultimate goal of the project is to allow our model to reconstruct high-resolution images from real-life low-resolution images, it is crucial that the downsampled images can well mimic real-life low-resolution images.

To alleviate this issue, the following future could be done in the next semester:

- Include various downsampling methods in the low-resolution images training data
- Add randomized blurring methods or noise to the low-resolution training data
- Save the low-resolution images into JPEG format. It is because JPEG uses an image compression method which could mimic images downloaded from the internet.

5.2. Difficulty in achieving state-of-the-art performance

Since the major objective of this project is to create super resolution models that could reconstruct the high-resolution image, the main challenge would be to achieve state-of-the-art performance in terms of PSNR and loss when evaluating on the evaluation data.
As mentioned above, the most recent and state-of-the-art face image super resolution models are very deep and consist of complicated architectures including FSRNET and SICNN. These models incorporate multiple features specific to face images like identity-wise loss, heatmap prediction network etc. In order to achieve a similar or better performance than these models, we hope to focus on building multi-branches deep learning models in which each branch of the model works on different tasks.

5.3. Difficulties of model evaluation

In this project, we focus on a very specific super resolution task, i.e. to super-resolve 27 x 22 face images 4 times larger to 108 x 88. Other existing models in the community are similar. So it is infeasible to directly compare the metrics as shown on the research paper as the scopes of the models are not the same.

Moreover, some of the super resolution projects don’t provide sample code to test the performance of the models. This makes it very difficult to reproduce the result as shown on the research papers and adjust the models in order to fit on our datasets.

In order to evaluate our model effectively and compare between our model and other existing model. We cloned the models provided with sample code like SRGAN into our working environment. So that we can train the model again on our training dataset and directly compare the result fairly with our models. For the projects that don’t provide sample codes, we have tried to implement models with the same architecture and training process in our environment.
6. Future Planning

6.1. Future development

The next step of the project involves the modification of existing implementations of our models so as to optimize their performance. It is projected that the project will try to incorporate GAN into some of our models to make the super-resolution image more realistic. While GAN is more difficult to converge, we will try to first train the SR model that incorporating face information, and only after this we can replace the simple loss function with the discriminative model.

Different network architectures may also be explored. In particular, locally connected layers could be useful when some features are functions of a small part of space, but each feature does not occur across all positions. For instance, when dealing with face image, the model only has to look for the eyes in the top half of the image. It is thus believed that locally connected layers could be useful for our purpose.

Since the hyperparameter tuning of different models is still incomplete, the next step of the project also involves optimizing the hyperparameters so as to optimize our algorithms. The loss functions could be better designed to take feature loss or hypersphere identity loss [28] into account. It is predicted that properly designed loss function could make the models better fit the project goal - to preserve the identity of the face and to produce accurate landmark locations.

6.2. Optional features implementation

After building a satisfactory model for face image super-resolution, we might extend our scope to relevant tasks if time allows. Such tasks include:

- Super-resolution of non-frontal face images
- MISR - face image super-resolution using multiple LR images of the same face
- Addition of visual effects to the upsampled face images
- Optimization on training and prediction speed.
- Face super-resolution in multiple video frames. This might be achieved by using a special class of neural network called Recurrent Neural Network.

6.3. Schedule

The current project progress is on schedule and is summarized in Table 6. The team is currently at stage 3 - model creation, evaluation and implementation.
Table 6: Project schedule

<table>
<thead>
<tr>
<th>Task</th>
<th>Estimated start time</th>
<th>Estimated completion time</th>
<th>Completion status</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stage 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initial self-study on deep learning</td>
<td>Aug 20, 2018</td>
<td>Oct 15, 2018</td>
<td>100%</td>
</tr>
<tr>
<td>Image collection and preprocessing</td>
<td>Sep 1, 2018</td>
<td>Sep 30, 2018</td>
<td>100%</td>
</tr>
<tr>
<td>Literature Review</td>
<td>Sep 1, 2018</td>
<td>Sep 30, 2018</td>
<td>100%</td>
</tr>
<tr>
<td>Phase 1 deliverable: Detailed project plan</td>
<td>Sep 1, 2018</td>
<td>Sep 30, 2018</td>
<td>100%</td>
</tr>
<tr>
<td>Phase 1 deliverable: Project webpage</td>
<td>Sep 1, 2018</td>
<td>Sep 30, 2018</td>
<td>100%</td>
</tr>
<tr>
<td><strong>Stage 2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Get familiar with tools and environment</td>
<td>Sep 15, 2018</td>
<td>Oct 15, 2018</td>
<td>100%</td>
</tr>
<tr>
<td>Replication of results from research papers</td>
<td>Oct 1, 2018</td>
<td>Oct 31, 2018</td>
<td>100%</td>
</tr>
<tr>
<td><strong>Stage 3</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model creation, evaluation and implementation</td>
<td>Nov 1, 2018</td>
<td>Jan 6, 2018</td>
<td>100%</td>
</tr>
<tr>
<td>● Deliverable: Deep learning model – initial version</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First presentation</td>
<td>Jan 7, 2019</td>
<td>Jan 11, 2019</td>
<td>100%</td>
</tr>
<tr>
<td>Phase 2 deliverable: Interim Report</td>
<td>Dec 24, 2018</td>
<td>Jan 20, 2019</td>
<td>100%</td>
</tr>
<tr>
<td>Enhance deep learning model</td>
<td>Jan 21, 2019</td>
<td>Feb 28, 2019</td>
<td>0%</td>
</tr>
<tr>
<td>● Deliverable: deep learning model – enhanced version</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Stage 4</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implement optional features</td>
<td>Mar 1, 2019</td>
<td>Apr 7, 2019</td>
<td>0%</td>
</tr>
<tr>
<td>● Deliverable: deep learning model – extended version</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phase 3 deliverable: Implementation and Final Report</td>
<td>Mar 18, 2019</td>
<td>Apr 14, 2019</td>
<td>0%</td>
</tr>
<tr>
<td>Final presentation</td>
<td>Apr 15, 2019</td>
<td>Apr 19, 2019</td>
<td>0%</td>
</tr>
<tr>
<td>Project exhibition</td>
<td>Apr 15, 2019</td>
<td>Apr 29, 2019</td>
<td>0%</td>
</tr>
<tr>
<td>Project competition (if selected)</td>
<td>Apr 30, 2019</td>
<td>May 29, 2019</td>
<td>0%</td>
</tr>
</tbody>
</table>
7. Conclusion

In the first semester, we set up the software environment for development. After reviewing relevant papers and codes, and collecting the face images, a complete workflow to train and evaluate the deep learning models has been created. The team started with replicating simple SR models, and continued with testing more complicated SR models. Next, various settings of the models and training process have been experimented.

At the next stage, face specific SR models have been built on top of the baseline models by incorporating face prior information like facial landmarks and recognition loss.

A general observation is the images produced by the models are often blurry which may be due to the objective function to be minimized. By continually developing models with face prior information, developing Generative Discriminative Network (GAN) and fine tuning the setting, we hope to produce sharper images and models with better performance. If time allows, we would implement other tasks related to SR.
References


