A Method for Face Image Inpainting Based on Generative Adversarial Networks

Xinyi Gao

A thesis submitted to the Auckland University of Technology

in partial fulfillment of the requirements for the degree of

Master of Computer and Information Sciences (MCIS)

2022

School of Engineering, Computer & Mathematical Sciences

Abstract

Recently, face image inpainting has become a fascinating research area in the field of deep learning. However, the existing methods have the disadvantage that the image inpainting results are not clear enough. Therefore, we propose a new face image inpainting method based on GAN (Generative Adversarial Network) in this thesis. Firstly, a deformation network based on GAN is designed. Then we add an identical autoencoder to the generative part of this generative adversarial network. Two loss functions of mean square error (MSE) loss and GAN loss are combined in the training process. Finally, through the analysis of results based on the CelebA dataset, the average of the new model's PSNR (Peak Signal-to-Noise Ratio) is 36.74dB, the average value of SSIM (Structural SIMilarity) is 0.91. Compared with the previous method, the new model has improved the effect of face image inpainting.

Keywords: Face Image Inpainting, Generative Adversarial Network, Convolutional Neural Network, Autoencoder

Table of Contents

List of Figures

List of Tables

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgments), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Signature: Date: 08 June 2022

Acknowledgment

Thanks to Shanxi Agricultural University China for giving me the opportunity to study in New Zealand.

I would like to express my sincere thanks to my main supervisor Wei Qi Yan. During the one and half year at the Auckland University of Technology, he offered me with professional technical guidance and support and pointed out the direction for me on the road of learning. I am very grateful to Dr. Yan for urging me to make continuous progress in my project. Also, I would like to thank my secondary supervisor Minh Nguyen for his academic support during my studies. Finally, I would like to thank the Auckland University of Technology for supporting students during the COVID-19 time, allowing me to continue my studies during the global epidemic.

Xinyi Gao

Auckland, New Zealand

June 2022

Chapter 1
Introduction

The first chapter of this thesis consists of five parts. In the first part, we introduce the background and motivation of this thesis and show that image inpainting has a wide range of applications that can be employed for accurate recognition of face images. By delving into the research background, in this chapter, we present the details of the research question and the specific contributions. The goals of this study are discussed in Section 1.4. Finally, we outline the structure of this thesis in Section 1.5.

1.1 Background and Motivation

As an essential branch of artificial intelligence, the field of computer vision (Rebecq, Ranftl, Koltun, & Scaramuzza, 2019) has been widely employed today. Computational vision refers to the processing of image and video data from various places by computers. Processing methods include identification, repair, etc. The main purpose of the field of computer vision is often to understand our human world through big visual data from digital cameras and help us to conduct data analysis. In this thesis, we will work on image restoration issues that have become increasingly important with the widespread use of digital cameras and smartphones.

During digital image processing, file corruption or data loss often occurs, which results in image corruption usually, images contain important information (Liu, Bai, Zeng, & Wang, 2019), but image restoration methods are often difficult to fix them quickly. Therefore, in order to improve the efficiency and repair accuracy, we need to find a new method to repair them quickly.

Image inpainting is a field of digital image processing in computer vision (Elharrouss, Almaadeed, Al-Maadeed, & Akbari, 2020). In short, its goal is to reproduce these images as realistic as possible, make use of the generated content to fill in missing or masked areas of the image, and utilize the inpainted image more realistic. Image inpainting is to use the information already in the image to repair the defective part of the image. It can be used on many occasions, such as cultural relics protection, film and television special effects production, restoration of old photos, removal of text in images, etc. (Liang, Jia, & Lu, 2021). Image inpainting was originally a traditional graphics problem, based on mathematics and physics. With the excellent results of deep learning in the field of vision, researchers began to turn their attention to deep learning, and image inpainting based on deep learning has become a research hotspot.

Traditional image inpainting methods are to fill the missing part by the pixels around the missing part of the image, the coarse method is to directly stretch the edge pixels, and the fine method needs to calculate the similarity of small patches. Image inpainting combined with digital techniques was firstly proposed in SIGRAPH 2000 (Bertalmio, Sapiro, Caselles, & Ballester, 2000). They took use of a texture-based algorithm to achieve the final inpainting by referencing both global and local images. This is the first application of deep learning methods to image inpainting. This idea refers to both global and local information, which improves the accuracy of inpainting. At that time, with the extensive development of information technology, which provided a good environment for the final realization of this technology, many people expected to integrate digital technology into image restoration. It turns out that image inpainting combined with deep learning is a good method.

Today, image inpainting has been widely applied to worn-out book restoration, medical image processing (Razzak, Naz, & Zaib, 2018) through Adobe Photoshop (C). Therefore, the research topic of image inpainting is worth studying. In addition, due to the complexity of imaging environment, there are a slew of difficulties in the restoration process. For example, the boundary between the original image and the inpainting location is visibly blurred in the inpainting result. In addition, how to ensure the global rationality of the inpainting area is one of the difficulties in the image inpainting task.

Nowadays, face image inpainting (Han & Wang, 2021) continues to be developed steadily. From the initial texture to the application of deep learning, face image inpainting has been developed rapidly. Recently, face image inpainting encounters these new problems. Most smartphones now are use of face unlock function, but due to the continued impact of the global epidemic, people generally wear masks to protect themselves while going out. Due to the occlusion of masks, if people need to unlock their mobile phones through face recognition or make automatic payments through face recognition, the existing methods cannot quickly unlock their mobile phones, which brings great inconvenience to our daily life. Therefore, most of researchers expect to apply a new face image inpainting method that allows people to use face recognition and unlock their phones while wearing masks. A fast and accurate face image inpainting method is urgently needed.

Due to various reasons, such as the maturity of image inpainting methods in the past, the existing demand for fast face image inpainting needs the in-depth knowledge from literature. Thus, we would like to ask a series of research questions for this thesis.

1.2 Research Questions

The purpose of this thesis is to study on how the process of face image inpainting is implemented and how these implementations can be improved. In this thesis, we fulfil face image inpainting through deep learning methods, and improve the accuracy and efficiency of the outcomes. Therefore, the research questions proposed in this thesis are:

(a) Whether the accuracy of face image inpainting can be improved by MSE (Qin, Chen, Shen, Jiang, & Feng, 2017) *loss and GAN loss* (Yu, 2019) *in GAN-based models?*

In the existing deep learning-based training models, most of them use the MSE loss function or the GAN loss function for training. In order to improve the accuracy of face image inpainting, we will firstly use the MSE loss function for training, then use the MSE loss function and the GAN loss function at the same time to train the model.

(b) Compared with existing GAN networks, how does the network proposed in this thesis improve the accuracy of face image inpainting in resource-constrained deep learning systems?

Modify the network model based on problem *(a)* and improve the accuracy of image inpainting by adding an autoencoder to the completion network.

The main purpose of this project is to implement face image inpainting by using deep learning methods. By modifying the network model and combining loss functions during training. Thereby improving the accuracy of face image inpainting.

1.3 Contributions

At the end of this project, we are able to achieve:

(a) *Image inpainting via GAN model.*

(b) *Improve the accuracy of image inpainting by jointly using MES loss and GAN loss function training during training.*

(c) *Improve the accuracy of face image restoration by adding an autoencoder to the completion network.*

In addition, we will compare the previous models and analyze the advantages and disadvantages of the two methods by repairing a diversity of parts of the face image.

1.4 Objectives of This Thesis

In this project, we firstly introduce related work on deep learning and face image inpainting. We will build a complete CNN (Convolutional Neural Network) based image inpainting network. Furthermore, we will use a GAN model in this image inpainting network to inpaint images. The GAN model mainly has a completion network and a discriminator network, and we will add an autoencoder to the discriminator network. Secondly, we will also jointly train the loss function. The two main loss functions in this project are MSE loss and GAN loss. Finally, in this thesis, we will analyze the results of our experiments by PSNR, SSIM and subjective judgment of human eyes.

1.5 Structure of This Thesis

The structure of the thesis is listed as follows:

In Chapter 2, a literature review will be presented, we will present past research work related to image inpainting. Firstly, we will introduce the basics of image inpainting and different approaches to apply deep learning to image inpainting processing. Finally, we will also review various studies on face image inpainting.

In Chapter 3, we depict the method proposed in this paper. In this section we describe the design and layout of the experiments. The datasets used for training and testing are introduced. In addition, methods for evaluating the results will be introduced.

In Chapter 4, we will implement algorithms and models. In this chapter, we will analyze and illustrate the experimental results in the form of pictures and tables. Here we will address limitations or issues of the thesis.

In Chapter 5, we analyze and discuss the results obtained in Chapter 4. Finally, we present conclusions and future work in Chapter 6.

Chapter 2
Literature Review

The focus of this project is on face image inpainting based on generative adversarial networks (GAN). In this chapter, we will introduce the process of development and related work of image inpainting.

2.1 Introduction

Today, most of face image inpainting methods are harnessed to repair those damaged images. This can help organizations or individuals that require full face information. As we mentioned in Chapter 1, with the continuous impact of global epidemic, in order to use the complete face information covered by masks more conveniently and quickly, it is necessary to repair face images with masks occluded. Therefore, researchers gradually shift the focus of attention to this area. In this thesis, we will conduct in-depth research investigation on the restoration of human face images.

Face image inpainting is booming with the emergence of various requirements related to inpainting (Sulam & Elad, 2016). With the increasing demands for face image restoration, such as restoration of damaged portrait artwork, destroyed suspect pictures, and spoiled commemorative photos that are of great significance to ordinary people. Face image restoration methods are gradually applied to art, photography, and a small number of security fields which received more and more attention.

The advent of deep learning and big data in computer science is key to develop new image inpainting methods. Deep learning is a part of machine learning (Dargan, Kumar, Ayyagari, & Kumar, 2020). Since its appearance, it has been applied to many fields such as computer vision, natural language processing, image analysis, etc. (Liu, et al., 2017). Image inpainting is to just the tip of the iceberg of deep learning applications. The main task of this thesis is to find a fast and convenient way to repair face images occluded by masks.

The way of image inpainting usually is to repair images from two aspects. One is global and the other is local. Firstly, the damaged image is initially repaired by referring to the local image, then we check from the global perspective whether the repaired image is reasonable. As a result, image inpainting methods demonstrate obvious research outcomes. In general, most of machine learning algorithms are based on the analysis of image data. The same is true in this thesis, where we are use of machine learning methods to inpaint and restore images.

2.2 Image Inpainting

At present, image inpainting is split into conventional methods and deep learning methods. The conventional image inpainting methods include sample-based texture synthesis methods, example-based structure synthesis methods, diffusion-based methods, sparse representation methods and hybrid methods (Jam, et al., 2020). The deep learning method repairs images through CNN and GAN.

2.2.1 Diffusion-Based Methods

During the Renaissance, a number of artists began to restore early works of art, the process was carried manually, the restoration process was subjective. However, the repaired results follow the overall structure of the image, and the repaired surrounding regions are all extended from the repaired area.

Bertalmin et al. proposed the BSCB image inpainting model in 2000 (Bertalmio, Sapiro, Caselles, & Ballester, 2000). The definition intended purpose and application classification of image inpainting are clearly presented. The purpose of image restoration is achieved by using a method based on partial differential equations. Its main idea is to iterate the information around the patched area into the patched area along the direction of iso-illuminance lines to generate patch information. The direction of the iso-illuminance lines can be calculated by calculating the discrete gradient vector of each point on the patched contour line and rotate it 90°. This effectively preserves the boundaries during iterations. After several rounds of iterations, the algorithm performs an iterative diffusion process to keep the patched area smooth. Anisotropic diffusion is beneficial for maintaining boundaries across the patched area. The algorithm has a good repair effect on creases and cracks. But it is only a simulation of the manual repairing process, and there is a slight blurring of the repaired edges.

Total variation is an image inpainting method based on the total variation model

(Rudin, Osher, & Fatemi, 1992). Its essence is to treat the image as a piecewise smooth function and build a model on the image in a bounded space. When this method is implemented, its partial differential equation should be digitized and converted into a difference equation to solve the partial differential equation through iterative operation. It was first used to denoise the image contaminated by noise, and then it was applied to image restoration.

A unified inpainting model was established based on the principle of energy minimization (Chan & Shen, 2000), which was applied to the field of image inpainting, and achieved good results. Moreover, due to the characteristics of the TV model, the more non-damaged pixels around the pixels to be repaired, the faster the information diffusion, the faster the repairing speed. Therefore, a new algorithm, the fast TV model (Lu, Wang, & Zhuoma, 2010), was developed, the damaged region of the image is splitted into multiple layers and repaired layer by layer. Since then, a great deal of improved algorithms based on the TV model have been proposed, such as a fast solution algorithm based on Alternating Direction Multiplier Method (ADMM) for processing non-smooth term convex optimization problems (He, Hu, Zhang, & Shi, 2014).

The image inpainting algorithm was utilized by combining net function interpolation and TV model (Lee, Lee, Kim, Cho, & Cho, 2018) et al. The most well-known among these improved algorithms is the Curvature-Driven Diffusion (CDD) model (Chan & Shen, 2001), which is proposed to solve the problem of edge fracture caused by the diffusion intensity in the TV model only depending on the iso-irradiation intensity. Therefore, the image curvature is added to the TV model. The diffusion is enhanced where the curvature is large, the diffusion is weakened where the curvature is small, so the problem of edge fracture can be well solved.

Oliveira model was proposed based on the existing models (Richard & Chang, 2001). The edge information of the damaged area is diffused to the damaged area through repeated convolution of a template with a fixed size, anisotropic diffusion is performed at the edge to prevent edge blurring, simple and fast to repair, and achieved good repair results.

2.2.2 Sample-based texture synthesis methods

Image inpainting with sample-based texture synthesis method is an algorithm that takes use of the similarity between blocks of the same image and images to be restored. The Criminisi algorithm (Criminisi, Perez, & Toyama, 2003), which is an image inpainting algorithm based on sample blocks, can well preserve the image structure while filling the defect region of the image, it has been verisied as a widely used and effective algorithm in image inpainting. The most important thing in the Criminisi algorithm is the priority calculation. The image restoration sequence is controlled by the priority, the optimal matching block is determined by the restoration priority of the block to be restored and the sample block.

Compared with partial differential equation based inpainting methods, the sample-based texture synthesis inpainting method utilizes more information of the complete area of the image. So, it can repair a large area of damaged images very well. But this method spends a lot of time in finding the best matching block and does not consider the local features of the image.

2.2.3 Methods based on sparse representation theory

As an efficient image representation method, image inpainting algorithms based on sparse representation are widely applied. Mallat firstly proposed the basic idea of image sparse representation (Yu, Sapiro, & Mallat, 2010), who employed an over-complete Gabor dictionary to sparsely represent images and proposed the Matching Pursuit (MP) algorithm. Neff et al. proposed a video coding algorithm based on Gabor dictionary and MP algorithm (Neff & Zakhor, 2002). Chen et al. proposed the Basis Pursuit (BP) algorithm (Chen, Donoho, & Saunders, 2001) to solve the convex optimization problem of l_1 norm in the sparse coding process. Guleryuz took use of the adaptive coefficient reconstruction method to estimate the optimal solution of the image repair area, and theoretically this method can obtain the local optimal solution.

The existing image reconstruction methods are required to pay special attention to the resolution of target images. In order to solve this problem, Zeng, et al. proposed a high-quality image reconstruction method (Zeng, Fu, Chao, & Guo, 2019), which is called Pyramid-context Encoder Network (PEN-Net). In the method, the structure of context encoder is increased, pyramid context encoder, multiscale decoder, and adversarial training loss are established. In this method, U-Net was employed as the backbone, the pyramid context encoder is applied to gradually fill in the missing content and ensure the consistency of the visual effects of image reconstruction. Thus, a multiscale decoder with deep supervision function is harnessed to calculate the loss. The use of this method allows the process to converge quickly during training time, a large number of experiments have proved the reliability and excellent performance of the proposed net.

After literature review, we find that deep learning method is widely employed in image inpainting. Thus, we also take use of deep learning methods to human face images. Therefore, in the following sections, we will introduce and explain deep learning methods furthermore.

2.3 Convolutional Neural Network

With the development of deep learning, multiple methods for human face image reconstruction are becoming more and more mature. A number of methods and frameworks for face image inpainting are emerging gradually. The typical methods include CNNs, which are available to attain higher-quality results in image inpainting.

CNNs are one kind of the important methods in deep learning. Although CNN can generate a reasonable structure in image restoration, the image generated by CNNs has structural inconsistency or fuzzy texture in the relevant regions. A new method (Yu, 2018) was proposed, the reason for this problem is identified, especially, when the deep net borrows textures from the surrounding areas.

Therefore, the method was derived from a generative model based on traditional texture and patch synthesis. This model is essentially a feedforward, fully connected neural network. The net can synthesize new image structures during inpainting, it has been verified to better use surrounding images as references. The experiments have proved that the model is effective to repair images from multiple datasets including human faces. The results are with higher quality than those existing methods. Later, a new system was proposed to learn from millions of images. The basic principle of this system is based on gated convolution, which eliminates extra marks. In the specific operation, partial convolution is summarized by providing a dynamic feature selection mechanism for each channel of spatial position of all layers.

As a method of deep learning, convolutional neural network has a slew of advantages in digital image processing. For example, using a neural network allows images to be directly fed into the network as input, because convolutional neural networks can reduce the complexity of the entire network model. Furthermore, since images can be directly fed into the network, this further avoids the problems of complex feature extraction and data reconstruction processes common with multiple image inputs (Albawi, Mohammed, and Al-Zawi, 2017). Therefore, Convolutional Neural Networks are gaining popularity in the image domain.

In fact, we treat convolutional networks as multilayer perceptrons. The input layer, convolution layer, sampling layer, fully connected layer and output layer constitute a basic typical CNN structure. In a typical CNN, convolutional and sampling layers alternate in the first few layers of the entire network. The order of composition is one layer of convolution layer, one layer of sampling layer, one layer of convolution layer, one layer of sampling layer and so on.

While the number of convolutional layers and sampling layers reaches the standard, the next step is the output layer of the fully connected layer, as shown in Figure 2.1. In the whole convolution process, since each neuron of the output feature surface in the convolution layer is locally connected with the input of the convolution layer, then the corresponding connection weights and the local input are weighted and summed together. The upper bias value is obtained to obtain the input value of the neuron. This process is equivalent to the convolutional neural network because of the convolution process, hence the name. Therefore, the key to convolutional neural network lies in the network structure, deconvolution and hole convolution, etc. In this section, we will firstly introduce the network structure and backpropagation algorithm of a simple CNN, and then give an overview of other commonly used CNN network structures and methods.

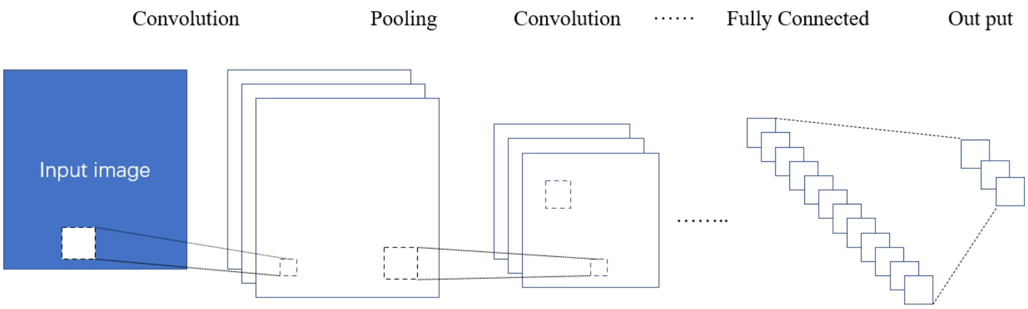

Figure 2.1: Convolutional neural networks

2.3.1 Convolutional layer

The convolutional layer of CNN is the part that extracts different features of the input by performing convolution operations. In the whole network, the initial convolutional layer often can only extract low-level features, such as lines, corners, etc. With the deepening of the number of convolutional layers, high-level convolutional layers can extract high-level features. Multiple neurons form a feature map, multiple feature maps form a convolutional layer. In other words, a convolutional layer is made up of many individual neurons. Neurons often exist as the basic processing units that make up artificial neural networks. The convolution kernel is a weight matrix. In the convolutional layer, each feature map is composed of multiple neurons, each neuron uses a convolution kernel to connect the local feature map in the previous convolutional layer and passes multiple inputs in the local area through the function Converted to a single output. (Silver, et al., 2016).

In order to get an output feature map, it is necessary to convolve the feature map of the previous layer with a convolution kernel. The convolution is activated through the activation function to obtain an output feature map. It is worth noting that one output feature map can often combine the convolution values of many different feature maps. In addition, the convolution kernel participating in the convolution must be a learnable convolution kernel (Wang, Jiang, Wang, & Wei, 2021):

$$x_j^l = f\left(u_j^l\right) \tag{2.1}$$

$$u_j^l \ = \sum\nolimits_{i \in M_j} x_i^{l-1} * k_{ij}^l + b_j^l \tag{2.2}$$

where u_j^l is called the net activation of the j^{th} channel of the convolutional layer l, which is obtained by using convolutional summation and bias of the output feature map x_i^{l-1} of the previous layer, x_j^l is the convolutional layer. The output of the j^{th} channel of $f(\cdot)$ is called activation function, usually functions such as sigmoid and tanh are harnessed. M_j represents the subset of input feature maps used to calculate u_j^l, k_{ij}^l is the convolution kernel matrix, b_j^l is the bias to the feature map after convolution. For an output feature map x_j^l, the convolution kernel k_{ij}^l corresponding to each input feature map x_i^{l-1} may be different, and * is the convolution symbol.

2.3.2 Pooling layer

The composition of the sampling layer (pooling layer) is similar to that of the convolutional layer, which is also composed of many feature surfaces to form a network layer. It is always working after the convolutional layer in the overall structure of CNN. Therefore, the input to the sampling layer always comes from the convolutional layer. The feature surfaces of the sampling layer correspond one-to-one with the feature surfaces of the input convolution layer, each feature map has a corresponding feature surface. After the input enters the sampling layer, it will not affect the number of feature surfaces as shown in Figure 2.1. In addition, neurons in the sampling layer are often also connected to part of the receptive field of the input convolutional layer and the receptive fields locally do not overlap. In fact, the purpose of the sampling layer is to reduce the resolution of the feature surface and obtain spatially invariant features by changing the resolution (Gu, et al., 2018).

The sampling layer will play the role of secondary feature extraction in the convolutional neural network. In addition, the sampling layer pools the local receptive field, which relies on neurons. Maximum pooling, taking the point with the largest value in the local receptive field, mean pooling, averaging all values in the local receptive field, random pooling, etc. are valuable methods (Boureau, Le Roux, Bach, Ponce, and LeCun, 2011). Among them, different pooling methods have different applications. For example, the max pooling method is suitable for separating some very sparse features (Boureau, Ponce, & LeCun, A theoretical analysis of feature pooling in visual recognition, 2010).

The sampling layer will firstly sample all input feature maps, then perform operations through the following formula, and finally obtain the output feature map:

$$x_j^l = f\left(u_j^l\right) \tag{2,1}$$

$$u_j^l \ = \beta_j^l down(x_i^{l-1}) + b_j^l \tag{2.2}$$

where u_j^l is called the net activation of the j^{th} channel of the down sampling layer *l*, which is obtained by sampling, weighting and biasing the output feature map x_i^{l-1} of the previous layer, β is the weight coefficient of the downsampling layer, and b_j^l is the downsampling layer. The bias term of the sampling layer. The symbol *down*(·) represents the down-sampling function, which divides the input feature map x_i^{l-1} into multiple non-overlapping *n* × *n* image patches by sliding window method, divides each image patch into multiple non-overlapping *n* × *n* image patches. The pixels are summed, averaged, or maxed out, so the output image is n-folded in both dimensions.

2.3.3 Fully connected layers

The basic structure of the fully connected layer is slightly different from the first two. Although the fully connected layer is still composed of neurons, each neuron in the layer is connected to the neurons of the previous layer in the network. As a neural layer following the convolutional layer and the sampling layer, the fully connected layer can integrate the feature information in the convolutional layer and the sampling layer, and the integrated information will have a certain distinction in categories.

In addition, in the CNN structure, the number of fully connected layers following the convolutional layer and the sampling layer is not fixed, it can be one fully connected layer or multiple fully connected layers (Sainath, Mohamed, Kingsbury, & Ramabhadran, 2013). The activation function in the fully connected layer often adopts the ReLU function, which should be used to improve the performance of the entire CNN network. And the activation function of all neurons in the fully connected layer is the same (Schmidt-Hieber, 2020). It is worth noting that in CNN networks, at the end of multiple fully connected layers, the output value of the fully connected layer is sent to a special output layer. This output layer is called SoftMax layer. This layer helps the CNN conduct simple classification by using SoftMax regression. This is the reason why the output layer is called a softmax layer. In addition, it is very important to choose a suitable loss function in CNN. For specific classification tasks, the wrong loss function often causes some unnecessary losses.

The input of the fully connected layer is often one-dimensional features. If the input is a two-dimensional feature, the input two-dimensional graphic features must be converted into corresponding one-dimensional image features and input into the fully connected layer. The output of the fully connected layer is obtained,

$$x^l = f(u^l) \tag{2.3}$$

and

$$u^l = w^l x^{l-1} + b^l \tag{2.4}$$

Furthermore, while training large feedforward neural networks, experiments cannot be conducted on small datasets. This is because feedforward neural networks tend to have a large capacity and, when tested with small datasets, often perform poorly on the held-out test data (also known as the validation set) (Yoo, 2015).

Additionally, there are still caveats to be aware of if an appropriate dataset is chosen for training. For example, in order to avoid overfitting during training, dropout techniques are often used in fully connected layers. This method is a regularization method. By using this method, the neuron nodes in the hidden layers of the neural network are disabled. The specific operation steps are to change the entire output probability of the neuron from 50% to 0, and achieve the result of invalidating the neuron. These non-working nodes will no longer participate in the operation of the CNN, that is, they will neither participate in the forward propagation process nor the back propagation process (Krizhevsky, Sutskever, & Hinton, 2012).

However, with the widespread use of dropout operations, this method still has certain defects. Therefore, a new technique is proposed based on dropout. This technique is called ReLU + dropout technique. This method solves the excessive randomness of the original dropout technology. The existing methods reduce the complexity between neuronal structures, make them more adaptive to each other. The improved technique makes the neurons in the network more robust. After experimentation, existing techniques achieve decent performance on many datasets. This has led to the fact that most people currently use ReLU + dropout technology when studying CNN (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) (Sainath, et al., 2015)

2.3.4 Deconvolution

The Deconvolutional networks model proposed by Zeiler et al (Zeiler, Krishnan, Taylor,

& Fergus, 2010). The idea is similar to CNN as a whole, but the direction is different. As an improved algorithm of CNN. In simple terms, deconvolution is the analysis of each layer in the CNN network by performing an inverse process. Throughout the inverse mapping of the eigenvalues of the convolutional network, the obtained eigenvalues are returned as input information or images again, which is the inverse process of the entire convolutional network. Therefore, the deconvolution network still belongs to a convolution model in essence, which has the convolution and pooling operations that the general convolution model has. But these operations are done in reverse.

After the concept of deconvolutional networks was proposed, a large number of studies followed. The most prominent of these is the applications. In the method, the deconvolutional network visualizes the convolutional features of the entire network as a complement to the convolutional network by adding a deconvolutional layer to all convolutional layers in the convolutional model. In this network, the output result of each convolution layer is not only used as the input value of the next layer, but also the output result is sent to the deconvolution layer for feature reconstruction, and the reconstructed result is compared with the original one.

After comparing the structure, the regions can be improved. The features visualized by the deconvolution network become very intuitive, we easily to see the improvements in the entire CNN network structure through these visual images (Zeiler & Fergus, 2014). By visualizing each layer of the CNN, Zeiler et al. came to the following conclusions. While rotating the features obtained in the CNN network, the features are greatly affected. After the feature is translated and scaled, the feature is not affected.

2.3.5 Dilated convolutions

The dilated convolution (Yang, Hu, Salakhutdinov, & Berg-Kirkpatrick, 2017) was originally proposed to solve the problem of image segmentation. The image segmentation algorithms usually take use of pooling layers and convolution layers to increase the receptive field (Receptive Field), which reduces the feature map size (resolution), and utilities upsampling to restore the images. In this process, the process of reducing and re-enlarging the feature map causes a loss of accuracy.

Therefore, an operation is needed that can increase the receptive field while keeping the size of the feature map unchanged to replace the downsampling and upsampling operations. It is the dilated convolution. Unlike normal convolutions, dilated convolutions introduce a hyper-parameter called dilation rate, which defines the spacing of values when the kernel processes the data. The expansion rate is also called the number of holes (Hole Size).

Hereinafter, we take 3 × 3 convolutions as an example to show the difference between ordinary convolution and dilated convolution, as shown in Figure 2.2, Fig. 2.3, Fig. 2.4.

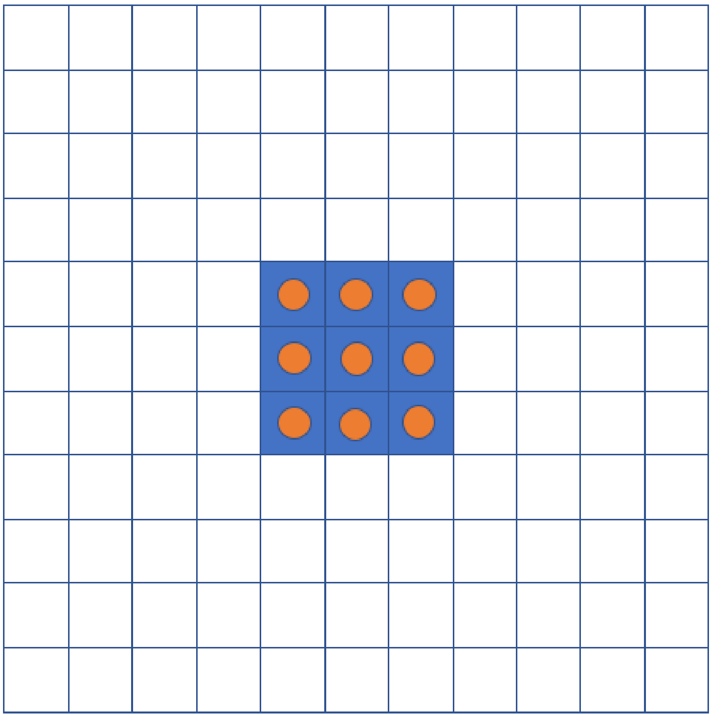

Figure 2.2: A 3×3 hole (Type I)

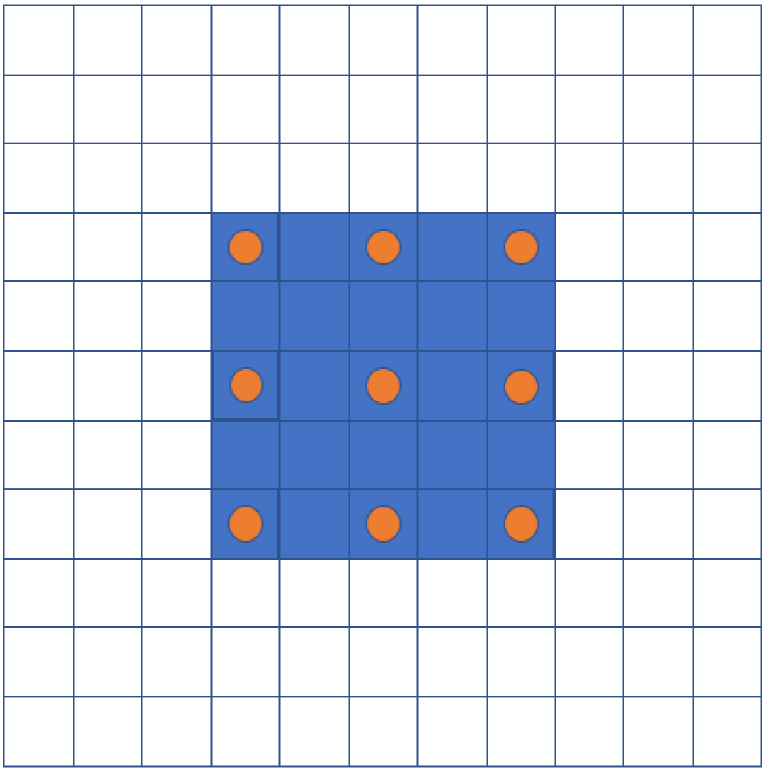

Figure 2.3: A 3×3 hole (Type II)

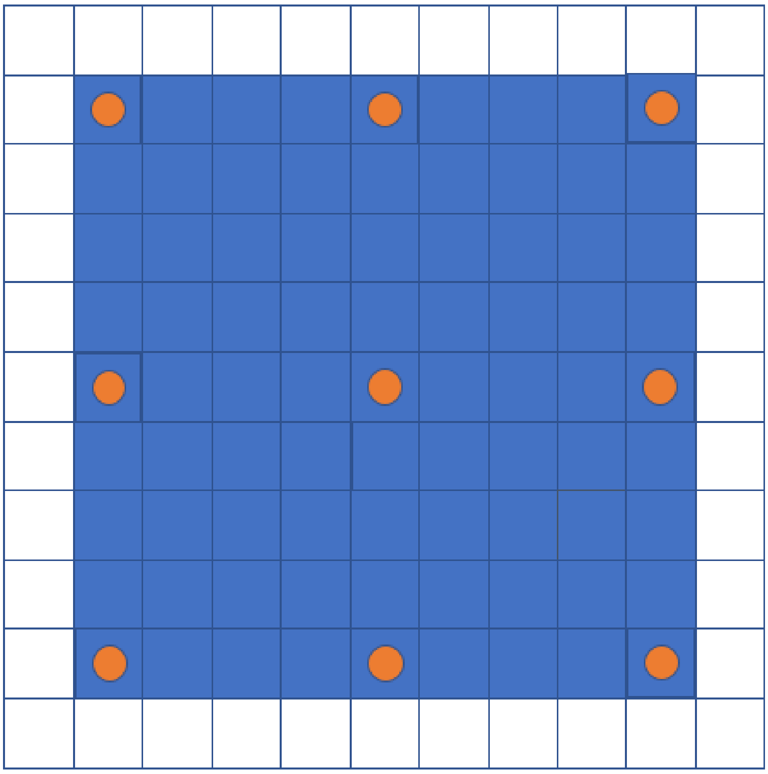

Figure 2.4: A 3×3 hole (Type III)

2.3.6 Feature surface

As an indispensable part of entire CNN, the number of feature surfaces is not constant (Wang, Patel, & Hacihaliloglu, 2018). But its number has a huge impact on the entire network. If the number of feature surfaces constituting the network is small, the network cannot effectively extract the features that are useful to the entire network. If the number of features is too large, the training time of the entire network will be greatly increased, affecting the efficiency of the network. Therefore, we set the number of feature surfaces, it needs to be set according to the actual needs of the network. If the number of feature faces is appropriate, it will not have a big impact on the network model.

2.3.7 Application in face image

CNNs can also be employed in combination with other methods, which have a great effect on solving specific problems. In order to solve the problem of blurred or missing face images due to acquisition method in the process of face recognition, conventional models for face image restoration often solve this problem from image viewpoint. The classic structure-based image restoration methods are CNN and generative adversarial networks. A face image reconstruction method was put forward based on generative confrontation network from a new perspective (Wei, Li, Liu, Zhang, & Chen, 2020). This method locates plane position of a face by determining two parallel lines of a vector. The different planes of the face are determined according to the given parallel lines, the edge curve is fitted through straight line segment to make facial contour clearer and make final facial features quite obvious.

Compared with the previous structure-based methods, this method can achieve better visual effects based on edges. The performance of face inpainting is greatly improved. It is worth mentioning that this method is only suitable for small-scale reconstruction processes. If there are too many missing regions in an image, the restoration result will become vague from the original image.

2.4 Generative Adversarial Networks

The recently popular generative adversarial network model GAN is one of future development directions of deep learning. Although Goodfellow et al. proposed GAN (Goodfellow, et al., 2014) in 2014, it was not until 2016 that researchers discovered the great potential of GAN. GAN has taken the breakthrough of bottleneck that previously limited the development of deep learning (Yu, Zhang, Wang, & Yu, 2017).

GAN is an emerging net for semi-supervised and unsupervised learning. It was firstly proposed in 2014, the entire adversarial process was achieved by implicitly modeling the high-dimensional distribution of the data. The main feature of this network is to train a pair of competing networks. GAN consists of two subnetworks, one is discriminative network, the other is generative network. The discriminative network is employed for both training and testing, but the generative network can only be employed for testing. The generative network is regarded as a counterfeiter, the other judgment network is regarded as an authenticator.

Forensic experts received fake and real images and aim to distinguish them. GAN pits two networks (i.e., generating network G and discriminative network D) against each other, both trained at the same time, and competing against each other. G continuously captures the probability distribution of real samples in the training set, and turns it into a fake by adding random noise. D observes real samples and fakes, and judges which is real, and which is fake (Creswell, et al., 2018). A simple network structure is shown in Figure 2.5.

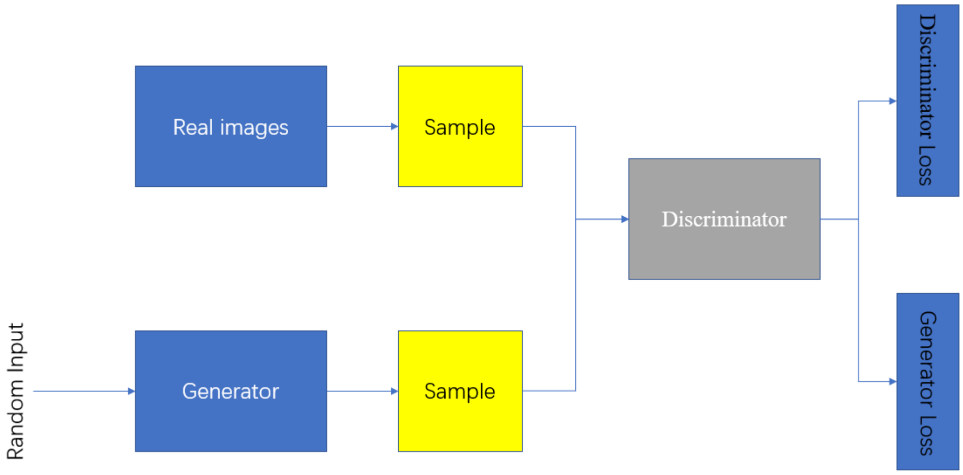

Figure 2.5: A simple GAN network

The most important part of this structure is that the generator has not direct relationship with the real image. Its only way is to learn how to grow through interaction with the discriminator. In fact, during the first training, the discriminator knows whether the image is from the real stack or from the generator that gets the error signal by accessing the synthetic image samples and the samples drawn from the real image stack, transferring these errors from the real image stack or from the generator to the generator. With the discriminator, the same erratic signals can be employed to train the generator and produce better quality fakes. By looping this process, the generated images become more and more realistic. This learning optimization process is to find a Nash equilibrium between the two.

Networks represented generators and discriminators are usually implemented by multilayer networks, which consist of convolutional and/or fully connected layers. The generator and discriminator networks must be differentiable functions, though they are not necessarily directly invertible.

Therefore, the discriminator and the generator are two indispensable parts in GAN. In order to analyze the whole GAN more intuitively, we take use of the function D to represent the discriminator and the function G to represent the generator. The two networks each have different missions. The discriminator D needs to judge whether the data it gets is real data x or generated fake data z. The purpose of generator G is to generate fake data that is as realistic as possible (Jia, Zheng, & Sun, 2019). The generator G wants the discriminator D to think that the fake data it outputs is real data, and the discriminator needs to distinguish the real and fake outputs through constant confrontation between the two networks, the two networks are continuously iteratively optimized, so that the entire network gets promote.

Finally, if the output of the discriminator D becomes stable, we think that the network has reached the optimal performance (Creswell, et al., 2018). If the GAN is directly modeled, the following objective function can be obtained,

$$f(x, z) = \min_{G} \max_{D}((D(x) - D(G(z)))). \tag{2.4}$$

2.4.1 GAN architectures

Due to the particularity of GAN network, GAN can continuously analyze the composition of real samples, and obtain the real structure of samples in constant confrontation. GAN network performs well in prediction and shows strong prediction; Compared with the samples generated by other machine models, the samples generated by the GAN network are more robust; because the GAN network is designed to be closer to the way humans think, the network can help artificial intelligence to be closer to humans and complete Some more complex tasks; in addition, GAN networks do not need to design a complex loss function. These are the advantages of GANs.

While GAN has many advantages, it also has a disadvantage, that is, the network is difficult to converge. In response to this problem, various GAN architectures have emerged. These different architectures that exist inside GANs are often applicable to different domains.

Goodfellow et al. created the generator and discriminator of GAN using multilayer perceptron. This is the first GAN to use a fully connected structure (Goodfellow, et al., 2014). This GAN model structure is relatively simple, so it is often used in simple image datasets. For example, the MNIST dataset related to handwritten digits, the CIFAR-10 dataset related to natural images, etc.

Creating a GAN with CNN is a further development of the fully connected GAN. Convolutional neural networks have very superior performance in processing image data and are closely related to multilayer perceptron. Therefore, the GAN network composed of CNN has appeared (Creswell, et al., 2018).

Laplacian pyramid of adversarial networks (LAPGAN) (Denton, Chintala, & Fergus, 2015) is a network that is different from general GANs. The network is structured as a concatenated network to form a pyramid. By using the LaPlace pyramid, the process of gradually generating high-resolution images is achieved by using low-resolution images and adversarial networks as a method. Its advantage is that only the residual between the sample and the generated sample is considered each time. As we know, LAPGAN is similar to the residual network.

Additionally, Radford et al. (Radford, Metz, & Chintala, 2015) proposed a network architecture called Deep Convolutional GAN (DCGAN) by training a generator and discriminator network consisting of deep convolutions. By training on three image datasets, the proposed network shows particularly good results, which has the potential in unsupervised learning. At the same time, the network also has problems. If training time of the model becomes longer and longer, the constructed network occasionally converts a subset of filters into other modes.

Adversarial autoencoder is a network of encoders and decoders with a structure similar to a GAN. In principle, after the image has been transformed by the encoder and decoder, it is restored and reconstructed again. The reconstructed image should be as similar as possible to the original image. This is also close to the principle of GAN. In other words, autoencoders can be viewed as a variant of GAN images in a way. Therefore, it also becomes logical to combine these two networks and use them in the image domain. Bengio (Bengio, Yao, Alain, & Vincent, 2013) provides a good example. They harnessed a generalized denoising autoencoder as a generative model of the network to explore the hidden relationship between the training process and the underlying data and achieved promising results.

2.4.2 Applications of GANs

GAN network is widely employed not only for speech and language processing, such as generating dialogue, generating images from text, etc. It can also be utilized in image and vision, speech and language, and other fields.

GAN can generate images that are consistent with the distribution of real data. A typical application comes from Twitter. The discriminator and generator are represented by using a parameterized residual network (Ledig, et al., 2017).

GANs have also begun to be applied to generate autonomous driving scenes. Santana et al. (Santana & Hotz, 2016) propose to use GAN to generate some images that are consistent with the actual traffic scene to help the training of the model, and then use an RNN-based model to make predictions on the trained model. Finally, very satisfactory results are attained. This shows that GAN can be applied to autonomous driving. That is, using GAN to continuously generate more realistic scenes to optimize the automaton is a promising model.

Gou et al. (Gou, et al., 2017) proposed an approach to achieve human eye detection using simulated and real images as training samples. But they also encounter a huge problem. That is the large gap between the simulated image used in the experiment and the real image, till GAN-based method (called SimGAN) was proposed to solve this problem (Shrivastava, et al., 2017). The synthesis error is reduced by introducing a self-regularization term into the network. In addition, they were use of unlabeled real images to make the synthetic images more realistic, while locally using an adversarial loss function to make the local information richer.

There exists related work in speech and language processing based on GANs. Li et al. (Li, et al., 2017) proposed to characterize the implicit correlation between dialogues by using GANs, thereby generating dialogue text. Zhang et al. (Zhang, Gan, & Carin, 2016) proposed GAN-based text generation, CNN is utilized as the discriminator, the discriminator takes use of moment matching to solve the optimization problem based on the output of the fitted LSTM.

In addition to apply GAN to the fields of image and vision, speech and language, there is the related work that integrates GANs and imitation learning (Ho & Ermon, 2016) (Finn, Christiano, Abbeel, & Levine, 2016), combines GAN and Actor-critic methods (Pfau & Vinyals, 2016), etc. MalGAN was proposed to detect malicious code that takes advantage of GANs to generate adversarial virus code samples (Hu & Tan, 2017). The experimental results show that GAN-based method can perform better than the traditional black-box or model-based methods.

GAN is also employed in the field of face recognition. The randomness of the mask was considered in the image, a GAN loss function was proposed, which is called SN-Patch GAN (Yu, et al, 2019). The experiments show that the results produced by the system have higher quality and much flexible results. This allows the system to assist users to quickly remove distracting objects, modify image layout, wipe off watermarks, etc.

In face recognition, the collection data is seriously distorted. A plenty of the collected face images are blurred or even lost. Faced with this problem, a 3D face image inpainting method based on generative adversarial network is proposed. The core of this method is to distinguish various plane structures by using two sets of parallel lines on different planes. According to the determination of different planes, the face features are extracted with three-dimensional. By fitting the edge curve, we make the facial contour clearer. In order to demonstrate the accuracy of experimental results, the results of previous experiments are compared. Compared with the past, the existing models can achieve visual matching while detecting edges in the missing regions of images. The performance of face recognition is significantly improved.

Image region missing is one of the most important losses in image damage at present. The existing image reconstruction algorithms still have shortcomings, such as blurry details and poor visual perception in terms of visual effects and algorithm efficiency after the reconstruction. In order to solve this problem, a new semantic restoration method (Zhang & Li, 2020) was proposed for facial images in 2020. On the basis of generating a confrontation network, the fusion of multiscale features of the given face to obtain more details without increasing the parameters. By expanding the receptive field in the deep net, the problem of insufficient edge information of the generated image is made up. In addition, learning ability of the generative network and the discriminant network is justified, which further improved the final performance of the proposed method.

Deep learning is the mainstream method for image restoration. The use of deep learning methods in image reconstruction can better restore the shape of human faces and obtain abstract features in the image. Therefore, in the same way, Han et al. (Han & Wang, 2021) also employed generative adversarial network as the basis for image inpainting process. However, a different method was proffered to solve vanishing gradient problem in the training process of GAN model. Evolutionary concepts were adopted to create a Generative Adversarial Network (EG-GAN) with an evolutionary generator for face image restoration. In the training process, EG-GAN updates the parameters of the generative network by combining two cost functions, generates offspring generators through crossover, and adds a matcher-assisted discriminator to criticize the generated images. Through the conception, the generative network continues to be evolved. This not only helps EG-GAN successfully overcome the vanishing gradient problem, but also improves the quality of image reconstruction and generated images that are in line with human vision.

GAN is an important model in the field of deep learning, which provides a powerful computational framework for unsupervised learning which is also an important tool for research in the field of images.

2.5 Multiple Networks for Face Image Inpainting

Combining multiple networks can often improve the result of image inpainting. Spatial-Temporal Nested GAN (STN-GAN) (Wu, Singh, & Kapoor, 2020) was proposed as a spatiotemporal information network. The model applies a specially trained image inpainting framework to video data by combining temporal information with residual blocks. This allows the network to repair not only the missing face images, but also the actual face images in the video. After the experiments on multiple public datasets, it is finally proved that the inpainting results of STN-GAN are spatiotemporally consistent. The final face restoration result is superior. Furthermore, a constrained inpainting method is proposed to restore the usability of damaged images.

Attention-based multistage generative networks for facial image inpainting are also used (Liu & Jung, 2021). This is because most face images lose a lot of content, resulting in blur and unnaturalness. Therefore, in this method, the number of channels of convolutional layers is reduced in the generator to improve the performance of inpainting. In addition, they combined attention with multilevel feature processing, by using multilevel feature processing to reduce training time while ensuring image relevance to surrounding content. Regarding the experiments, various loss functions namely mean absolute error (MAE), edge-preserving losses, adversarial and perceptual losses are adopted. Final experiments show that the proposed method produces superior inpainting results in random masks, outperforms the state-of-the-art methods during the same time.

In order to improve the quality of face inpainting, a domain-embedded generative adversarial network (DE-GAN) for face inpainting has been proposed (Zhang, et al., 2022). DE-GAN embeds face mask, face part, and landmark images into a latent variable space via a Hierarchical Variational Autoencoder (HVAE) to guide face completion. Compared with other GAN networks, this network is use of a global discriminator and a patch discriminator for judging whether the generated results are excellent. Experimental results show that this method produces higher quality inpainting results than contemporaneous methods and achieves state-of-the-art performance.

2.6 Autoencoder

The first use of autoencoders was in the 1980s as a feature extraction method. Nowadays, with the development of information technology, data such as various information and images are becoming increasingly high-dimensional. Before using machine learning methods, we often need various means to reduce the dimensionality of the dataset to reduce the computational difficulty. The autoencoder is one of the most popular methods for dimensionality reduction.

As a typical unsupervised neural network, the purpose of autoencoder is to express or map high-dimensional input images in a low-dimensional way. In simple terms, an autoencoder is the process of restoring the input data or image. We can briefly understand the construction of the autoencoder from Figure 2.6.

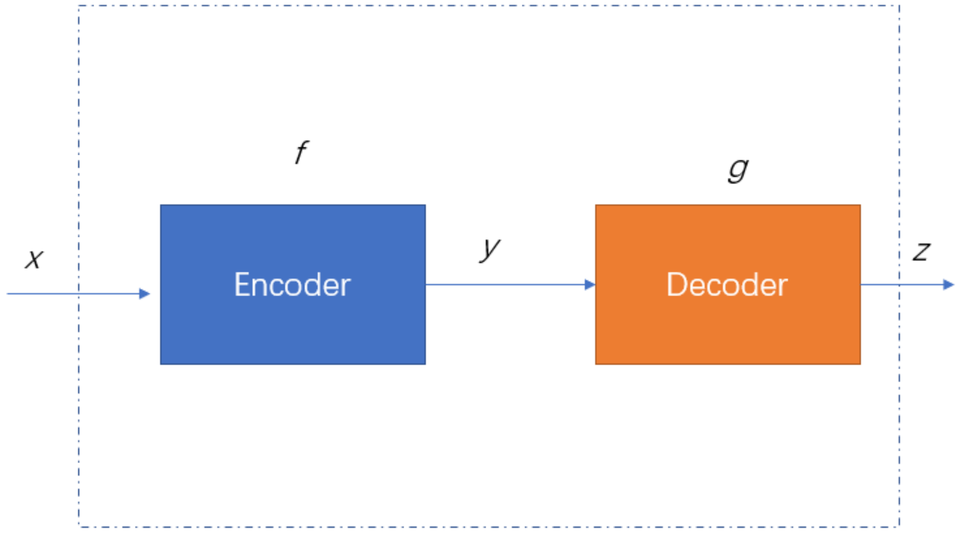

Figure 2.6: The structure of autoencoder

In Figure 2.6, the main purpose of the autoencoder is to convert the input x to an intermediate variable y, then convert y to z, then compare the input x and output z as well as make them infinitely close.

The data, after dimensionality reduction by using the autoencoder, often contains the main information of the original data. As a result, autoencoders tend not to make large changes if the original data changes slightly.

2.6.1 Architecture of autoencoders

Usually, an autoencoder consists of two parts: An encoder and a decoder. Among them, the encoder is a module that compresses the input data or images. The compressed data is often much smaller than the original data. The decoder acts as a module that reconstructs the encoded data. It is responsible for decompressing the compressed data and reconstructing it. Finally, the reconstructed output original data is compared to determine whether the decoding is successful.

An autoencoder needs to be trained as a neural network. There are four parameters that are crucial in training of an autoencoder, the encoded size, the number of layers of the autoencoder, the nodes of each layer, and the loss function. The encoded size can help the autoencoder to better decide the range of the data and help the autoencoder to regularize. The number of layers and nodes determines the processing speed of the autoencoder and the complexity of the model. The loss function helps the autoencoder determine the loss of the output.

2.6.2 Types of autoencoders

There are roughly five types of autoencoders, namely undercomplete autoencoders, sparse autoencoders, contractive autoencoders, denoising autoencoders, and variational autoencoders. The autoencoders have various applications.

Undercomplete autoencoders as one of the simple types of autoencoders. The structure of this autoencoder is very simple that is a true unsupervised encoder. The purpose of the encoder is to output the same image as received.

The purpose of sparse autoencoders is basically the same as that of undercomplete autoencoders, but the encoder is adjusted by using the number of nodes in the hidden layer. Using this encoder can help the neural network to regularize better.

Contractive autoencoders help autoencoders to better reconstruct results and resist the impact of input. Denoising autoencoders are employed to remove noise from images. There are often a lot of noises in the input image of the autoencoder, the autoencoder needs to remove the noise in the image throughout the process of encoding and decoding. Variational autoencoders are autoencoders that are trained to avoid overfitting. This is the reason why the autoencoder is often used to reconstruct the model.

2.6.3 Applications of Autoencoders

Since the emerging of autoencoders, the applications have become more and more in-depth, many related applications based on autoencoders have been developed.

The most popular application is semantic recognition using autoencoders. A method was propound to encode each polysemy by using a redesigned network (Liou, Cheng, Liou, & Liou, 2012). The proposed method has an automatic way to assign multiple codes to polysemy. This may benefit research in text mining, sorting, indexing, and classification. Polysemous words often have multiple meanings, which creates confusion for machines. Therefore, this method is proposed, which takes use of Elman network to process word sequences in literary works. The method has a wide range of applications, starting with the ranking, indexing and classifying literary works. In the work, the method was employed to the fashion analysis of two Chinese novels: “A Dream of Red Mansions” and “Romance of the Three Kingdoms.”

Besides semantic analysis, autoencoders have other applications. The applications are grouped as dimensionality reduction, image denoising, anomaly detection, and generation of time series. Autoencoding algorithms can also be employed for data dimensionality reduction. A novel dimensionality reduction method based on autoencoders (Wang, et al. 2014) has been developed. The method is implemented by using a "generalized autoencoder" (GAE). This method takes use of an iterative method to explore the relationship between data and form a manifold structure which is use of the relationship to pursue manifold structure.

In this approach, traditional autoencoders are extended in two ways. Furthermore, a multilayer architecture of generalized autoencoders, called deep generalized autoencoders was proposed to handle the datasets with complex conditions. After experimented on three datasets, the proposed method was demonstrated that had superior function. In fact, we consider the autoencoder as an unsupervised neural network whose architecture is perfect.

Therefore, we design our face image inpainting network based on the architecture of existing autoencoders. The specific structure and content will be introduced in the next chapter. Meng et al. proposed new applications for feature extraction of autoencoders. In the view from the encoder as a neural network-based feature extraction method, the success in abstracting features is worthy of recognition. However, the existing methods of autoencoder do not consider the existence of hidden relationships between data samples, this omission will affect the experimental results of the final autoencoder. Therefore, a relational autoencoder model was proposed with data features and the relationships. After demonstrated the feasibility of the method, the model was extended to other major autoencoder models, mainly including sparse autoencoders, denoising autoencoders and variational autoencoders.

In this section, we firstly introduce the models and methods that have been widely employed for image inpainting or face image inpainting. In addition, the concepts and applications of multiple deep learning methods, including CNN, GAN, and autoencoders, are introduced, all of which can be employed for face image inpainting. Both CNNs and GANs take a large number of training samples to get the best model and predict a given sample from the trained model. Therefore, all classes are known, the ambiguity of the training samples is low.

Autoencoders do not require training samples, which often extract more effective new features from neural network models or perform feature dimensionality reduction. So, the autoencoder can act as a feature extractor. In the following chapters of this these, we will illustrate the method and model we are using in this thesis.

Chapter 3
Methodology

In this Chapter, we present the research methodology in Section 3.3. In this chapter, we introduce deep learning methods for face image inpainting. The focus on this chapter is on the details of deep learning methods.

3.1 Convolutional Neural Networks

Convolutional neural network is a representative algorithm of deep learning, which includes convolution computations and deep structure. In recent years, with the improvement of deep learning theory, convolutional neural networks have been developed rapidly. At present, convolutional neural networks have been applied to image processing in computer vision. Figure 3.1 shows the basic CNN structure, which consists of an input layer, a hidden layer, and an output layer.

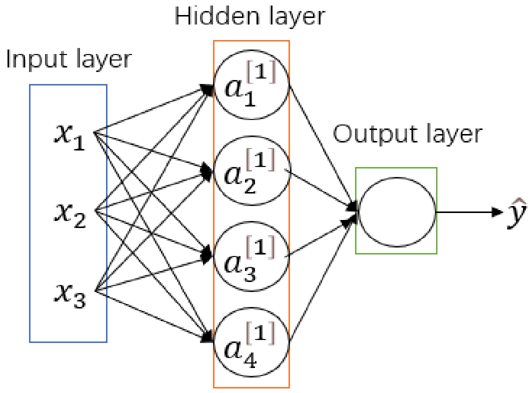

Figure 3.1: The basic structure of CNN

The CNN model is expressed as,

$$\hat{y} = \sigma(t), t = w^T + b \tag{3.1}$$

where $\sigma(\cdot)$ function is Rectified Linear Unit (ReLU) (Liu & Jung, 2019). This model is expanded as,

$$\begin{gathered}
\mathrm{z}_1^{[1]} = \mathrm{w}_1^{[1]\mathrm{T}} x + b_1^{[1]}, a_1^{[1]} = \sigma\left(\mathrm{z}_1^{[1]}\right) \\
\mathrm{z}_2^{[1]} = \mathrm{w}_2^{[1]\mathrm{T}} x + b_2^{[1]}, a_2^{[1]} = \sigma\left(\mathrm{z}_2^{[1]}\right) \\
\mathrm{z}_3^{[1]} = \mathrm{w}_3^{[1]\mathrm{T}} x + b_3^{[1]}, a_3^{[1]} = \sigma\left(\mathrm{z}_3^{[1]}\right) \\
\mathrm{z}_4^{[1]} = \mathrm{w}_4^{[1]\mathrm{T}} x + b_4^{[1]}, a_4^{[1]} = \sigma\left(\mathrm{z}_4^{[1]}\right)
\end{gathered} \tag{3.2}$$

where $a^{[n]}$ represents the matrix composed of the activation functions of the neurons in the n-th layer, $z^{[n]}$ function represents the matrix showing the processing logic of the neurons in the n-th layer. The activation function a generates a reasonable output through the data of z function which passed to the next layer.

In image inpainting, the input to the convolutional layer is usually a 2D channel,

$$I_{input} = H * W \tag{3.3}$$

where W represents the width of the image, H shows the height of the image, S indicates the step size of the convolution kernel, P (Padding) is the number of boundary pixel layers added to the edge of the image. The size of the image after passing through the convolutional layer,

$$W_{output} = \frac{W_{input}-W_{filter}+2P}{S} + 1 \tag{3.4}$$

and

$$H_{output} = \frac{H_{input}-H_{filter}+2P}{S} + 1 \tag{3.5}$$

These networks consist of convolutional layers that contain a set of filters. After each convolutional layer, in addition to the last layer, there is a ReLU layer. The input image is passed through a filter to generate an output image, which is then further processed using a nonlinear activation function. The most common nonlinear activation function is ReLU,

$$\sigma(x) = \max(0, x) \tag{3.6}$$

Linear rectification function is an activation function in convolutional neural network. This activation function defines the output of the neuron $w^T x + b$ after a linear change. The resulting output is nonlinear. For an input vector x, a neuron using a linearly rectified activation function output $\max(0, w^T x + b)$. The most popular CNN units are shown in the Eq. (3.7),

$$\hat{y} = f(\sum_i w_i x_i + b), f \in \{ReLU, sigmoid\} \tag{3.7}$$

The activation functions are sigmoid (Yin, Goudriaan, Lantinga, Vos, & Spiertz, 2003) function,

$$Sigmoid(x) = \frac{1}{1+e^{-x}} \tag{3.8}$$

The linear activation function simply sets the threshold as zero, reducing computational overhead. The output layer consists of a convolutional layer with a sigmoid function instead of a ReLU layer. The sigmoid function normalizes the output to the range [0, 1].

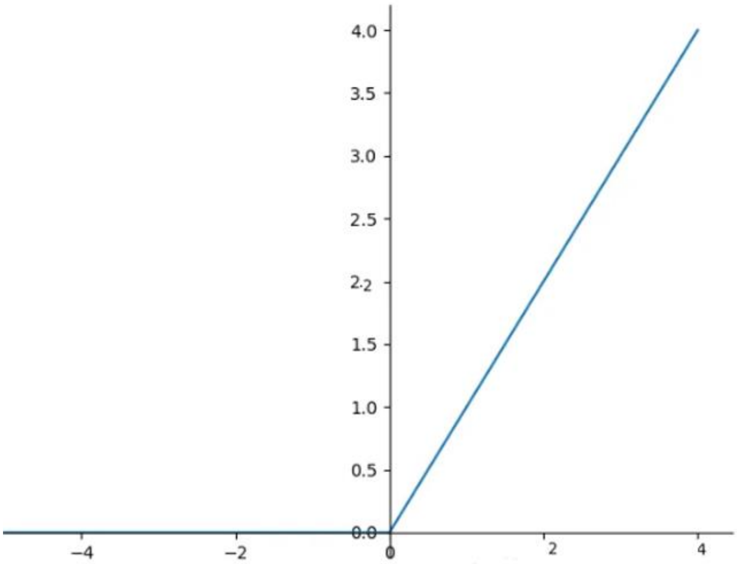

Figure 3.2: The curve of ReLU

In our proposed network structure, we are use of not only convolutional layers but also a variant called dilated convolutional layers. Dilated convolution is to inject holes in the standard convolution map to increase the reception field. Compared with the original normal convolution, the dilated convolution has one more hyperparameters called dilation rate, which refers to the number of intervals of the kernel.

Pooling can expand the receptive field and reduce the data dimension, but the information in the region will be lost. Regarding semantic segmentation, this creates a development bottleneck. The dilated convolution can expand the receptive field without losing information. So instead of adding pooling after each convolutional layer, we add dilated convolutions after multiple convolutions.

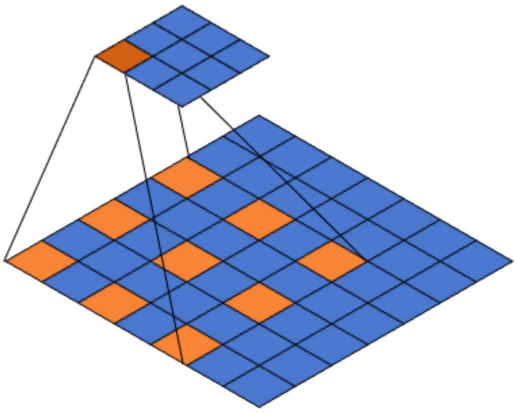

Figure 3.3: The process of dilated convolution

The RNN units are shown in Eq. (3.9),

$$h_t = f(W_x x_t + W_h h_{t-1} + b), f \in \{sigmoid, tanh\} \tag{3.9}$$

where input x_t and output h_t are vectors, W_x and W_h are the weigh matrices of x_t, h_{t-1}, respectively. The activation functions are sigmoid functions and tanh functions,

$$Sigmoid(x) = \frac{1}{1+e^{-x}} \tag{3.10}$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{3.11}$$

The sigmoid function maps the real numbers to the range, the tanh function maps the real numbers to the range, the images are shown in Figure 3.4.

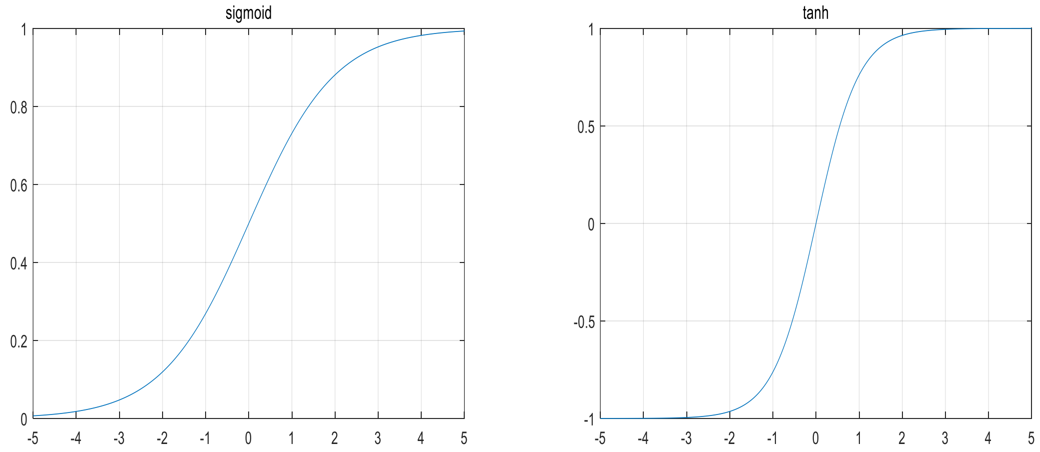

Figure 3.4: The *sigmoid* and *tanh* functions

The activation functions are to increase the expressiveness of the entire network. The tanh(·) function is the result of downward translation and scaling of sigmoid. In Figure 3.4, the effect of using the function tanh in the hidden layer is better than the sigmoid function. Afterwards, the convolution operation is conducted, which is an effective method for extracting image features. Generally, a square convolution kernel is applied to traverse each pixel in the image. Convolution is a local operation, the local information of the image is obtained by acting on the local image area with a size of convolution kernel. Since we have repaired the color face image in this thesis, we need to conduct cube convolution.

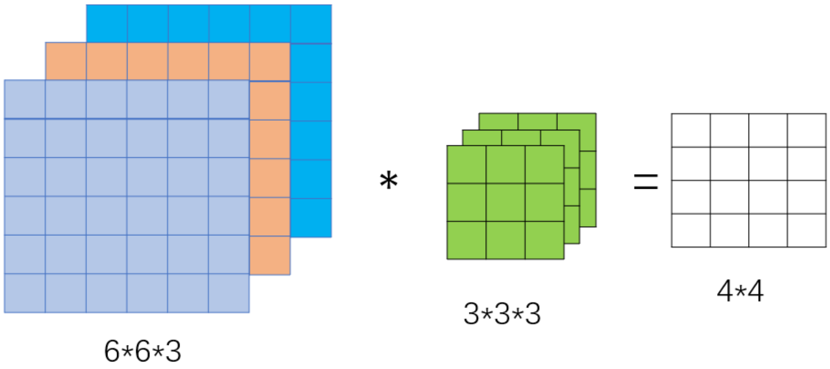

Figure 3.5: The cube of convolutions

The final step of convolution operation is to feed the obtained result into a fully connected layer (Liu, Kang, Zhang, & Hou, 2018). The main role of the fully connected layer is to map the features to the label space of samples. We get a one-dimensional vector in the final stage of training the network.

3.2 Autoencoder

Autoencoders are a type of unsupervised learning. The process of an autoencoder is to take an input, convert it into an efficient internal representation, and finally output an analog of the input. An autoencoder usually consists of two parts, an encoder and a decoder. The role of the encoder is to convert the input into an internal representation, the decoder converts the internal representation into the output. The numbers of input neurons and output neurons in the autoencoder are equal.

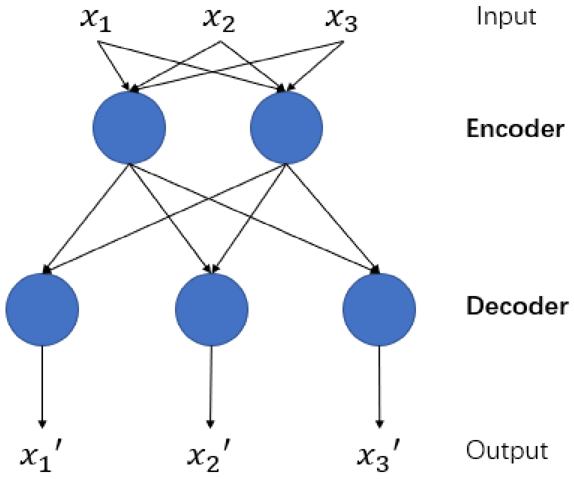

Figure 3.6: A simple autoencoder

3.3 Training Data

In this research project, we took use of the CelebA dataset with 202,599 face images for training (Liu, Luo, Wang, & Tang, 2018). CelebA is a large-scale dataset dedicated to face experiments, containing more than 200,000 face images. The image size in the CelebA dataset is all 178×218. In addition, the backgrounds of the various face images in this dataset are often complex, which make use of this dataset as a very suitable training dataset for this thesis. There are 5,000 face images which were randomly selected for training and testing.

In order to make the experimental results easier, we resize all image pixels to 128×128. During model training, we randomly add a mask image with a pixel size ranging from 24×24 to 48×48 so as to generate training data. In order to compare and analyze the experimental results, we also took use of the CelebA dataset for the test dataset, and randomly added a mask to it.

3.4 Research Designing

In this thesis, we utilize CNNs in deep learning as the base network, we are use of the GAN model in deep learning. As a typical method for face image reconstruction, GAN (Frid-Adar, Goldberger, & Greenspan, 2018) is also one of the most important methods in this thesis, which was firstly proposed in 2014. The basic idea of this method is to deploy two contrastive neural networks against each other. This is a way to get better results by playing against two networks.

GAN consists of a generative network and a discriminative network. The generator needs to generate more realistic images, the discriminator judges how “real” the input is. During training, weighted mean squared error (MSE) loss function and GAN loss function are used. They improved the stability of model training, the convolutional network of the whole network structure is employed for face image inpainting (Iizuka, Simo-Serra, & Ishikawa, 2017).

Furthermore, the global and local discriminator networks were also proposed to improve the quality of image inpainting. In fact, the ultimate purpose of complete network training is to fool the judgment of the discriminator network and let the recognition network output real images. The images were not reconstructed by using the completed network. Therefore, this can be achieved by simultaneously training all the networks during training, thereby improving the quality of the final output image.

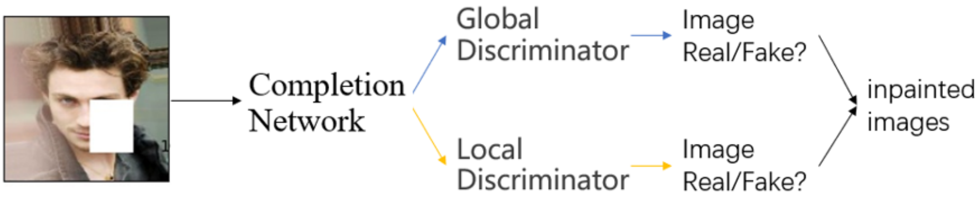

Figure 3.7: GAN training process

The GAN network is shown in Figure 3.7. This network consists of a generator and a discriminator. The discriminator part is composed of a global discriminator and a local discriminator. The discriminator is applied to judge whether the image is real or not.

In this thesis, we design two GAN-based image inpainting methods whose internal network consists of CNNs. The overall structure of the two methods is composed of a completion network and a discriminator network, the discriminator network is divided into a local discriminator and a global discriminator. Afterwards, the two discriminators are connected together through a fully connected layer to obtain the final inpainted image. The model for the first approach is called one-autoencoder GAN. The model for the second approach is called two-autoencoder GAN.

3.4.1 Completion Network

The one-autoencoder GAN model is mainly composed of convolutional neural network in the completion of the network. The convolutional layer is the core part of the whole network. The function of the convolution layer is to perform feature extraction on the input data, which contains multiple convolution kernels inside. Each convolution kernel is a three-dimensional digital matrix, the parameters of a set of convolution kernels include the height and width of the filter, the number of channels of the input image, and the number of convolution kernels. At the same time, the step size in the convolution kernel is the number of columns or rows that the convolution kernel slides each time, which will affect the pixel size of the output image, in order to facilitate the observation of the related operations of the convolutional neural network. The size of convolution stride determines the distance of the filter each time it scans the feature map (Wu, Li, Herencsar, Vo, & Lin, 2021). In this content, the default stride is one unless otherwise specified.

Generally, in a convolutional neural network, there is a pooling layer after the convolutional layer. The main function of the pooling layer is to reduce the dimension of features. An image contains many features, but when we identify an image, we often need the most prominent features, so we are use of pooling operation to reduce the dimension of the features. At the same time, pooling can also remove redundant information and retain valuable information.

There are two types of pooling, max pooling, and average pooling (Vargas, Esquivel, & Tickoo, 2021). Maximum pooling is to obtain the maximum value in the neighborhood, which can retain prominent features such as texture near the lost area. Average pooling is to average the pixels in the neighborhood, this method can retain basic information such as image background. Adding a pooling layer after multiple convolutional layers can effectively avoid overfitting.

We do not use pooling layers. In this project, we will replace the role of pooling layer by using dilated convolutions. Dilated convolution is a method to increase the receptive field. Dilated convolution adds holes to standard convolution and reduce the number of computations. In addition, we also harneness two deconvolutions in the completed network part. Deconvolution is the inverse process of convolution (Iizuka, Simo-Serra, & Ishikawa, 2017). Throughout deconvolution, the image can be converted into a 3-channel image again.

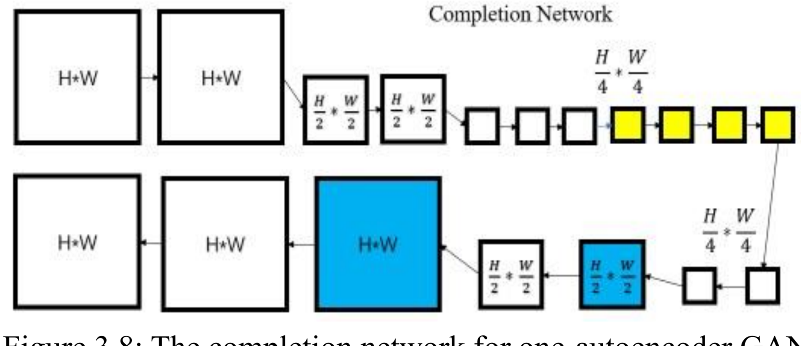
Completion Network
H*W
H*W
$\frac{H}{2}*\frac{W}{2}$
$\frac{H}{2}*\frac{W}{2}$
$\frac{H}{4}*\frac{W}{4}$
$\frac{H}{4}*\frac{W}{4}$
$\frac{H}{2}*\frac{W}{2}$
$\frac{H}{2}*\frac{W}{2}$
H*W
H*W
H*W

Figure 3.8: The completion network for one-autoencoder GAN

The completion network of the one-autoencoder GAN model consists of 12 layers of convolutional layers, 4 layers of dilated convolution layers and 2 layers of deconvolution layers. There are eighteen layers in total, where white represents convolutional layers, yellow represents dilated layers, and blue represents deconvolutional layers. The parameters of the completion network are shown in Table 3.1. The stride of convolutional layer 3 and convolutional layer 5 is 2×2.

Table 3.1: The completion network

Convolution types	**Parameters**			
	Kernels	***Dilations***	***Strides***	***Outputs***
Convolution layer1	5×5	None	1×1	64
Convolution layer2	3×3	None	1×1	64
Convolution layer3	3×3	None	2×2	128
Convolution layer4	3×3	None	1×1	128
Convolution layer5	3×3	None	2×2	256
Convolution layer6	3×3	None	1×1	256
Convolution layer7	3×3	None	1×1	256
Dilated layer1	3×3	2	1×1	256
Dilated layer2	3×3	4	1×1	256
Dilated layer3	3×3	8	1×1	256
Dilated layer4	3×3	16	1×1	256
Convolution layer8	3×3	None	1×1	256
Convolution layer9	3×3	None	1×1	256
Deconvolution layer1	4×4	None	$\frac{1}{2}\times\frac{1}{2}$	128
Convolution layer10	3×3	None	1×1	128

Convolution types	**Parameters**			
	Kernels	***Dilations***	***Strides***	***Outputs***
Deconvolution layer2	4×4	None	$\frac{1}{2}\times\frac{1}{2}$	64
Convolution layer11	3×3	None	1×1	32
Convolution layer12	3×3	None	1×1	3

3.4.2 Discriminator Network

The global discriminator and the local discriminator form the discriminator network. The role of the global discriminator is to consider the global image more in the process of image inpainting. The local discriminator pays much attention to the details of image inpainting. The global discriminator and the local discriminator consist of convolutional neural networks. Since the local discriminator consists of four convolutional layers with stride 2×2, whereas the global discriminator consists of 5 convolutional layers with stride 2×2.

The images for training were from the processed CelebA dataset. The image sizes in the dataset are all changed to 128×128. Due to the low resolution of the images, compared with Iizuka's network, we prune one layer of the global discriminator and local discriminator respectively. The discriminator evaluates samples from the training data. Both discriminators are designed to improve the realism of the output image. The results of the two discriminators are combined by a fully connected layer after passing through the discriminator network.

Table 3.2 The local discriminator

Convolution types	**Parameters**			
	Kernels	***Dilations***	***Strides***	***Outputs***
Convolution1	5×5	None	2×2	64
Convolution2	5×5	None	2×2	128

Convolution types	**Parameters**			
	Kernels	***Dilations***	***Strides***	***Outputs***
Convolution3	5×5	None	2×2	256
Convolution4	5×5	None	2×2	512

Table 3.3 The global discriminator

Convolution types	**Parameters**			
	Kernels	***Dilations***	***Strides***	***Outputs***
Convolution1	5×5	None	2×2	64
Convolution2	5×5	None	2×2	128
Convolution3	5×5	None	2×2	256
Convolution4	5×5	None	2×2	512
Convolution5	5×5	None	2×2	512

3.4.3 Algorithms

The loss function is an essential element in deep learning. It is generally applied to measure the degree of inconsistency between the predicted value of the model and the actual value. The loss function can give a neural network a lot of practical flexibility, it will define how the output of the network is connected to the rest of the network. Two loss functions are employed in our experiments. One of the loss functions is GAN loss and the other is MSE loss. During model training, we are use of a loss function that combines MSE loss and GAN loss.

MSE Loss

Mean Squared Error (MSE) is a popular loss function. The MSE function is applied to calculate the gap between the trained model and the actual model. The calculation result needs to take evaluation error between the predicted value and the actual value. You need to square the errors between the prediction and the ground truth (Yang, et al., 2021). Then, it averages over the entire dataset. Finally, the MSE loss values are obtained. The MSE loss function is shown as eq. (3.12).

$$L_{mse} = \|M_i \odot (C(I, M_i) - I)\|^2 \tag{3.12}$$

In the MSE loss function, $C(I, M_i)$ represents the completion network, I is the input image, $\odot$ is pixel-wise multiplication (Zhang, Quan, Wu, Li, & Yan, 2020). During the training, we mark the regions that need to be repaired as '1' and the complete regions of the image as '0'. The value of MSE loss is always greater than 1.00. The closer the value is to 1.00, the more realistic the training results.

We are use of MSE loss function to evaluate the model after each model training. After the network goes through multiple rounds of iterations, the output of the loss function gets closer and closer to 1.00. The performance of the network becomes more and more stable.

GAN Loss

A GAN can have two loss functions: One for generator training, the other for discriminator training (Chen, et al., 2021). Both loss functions follow the max-min loss decision rule and output the mathematical expectation as the loss. In this experiment, we mainly take use of the minimax loss in the GAN model. In the minimax loss function, $D(I, M_d)$ represents the discriminative network, $C(I, M_i)$ shows the completion network, M_d is a randomly generated mask, and I is the input image. M_i is the mask image with the exact dimensions of the input image. The GAN loss function formula is shown in eq. (3.13).

$$L_{GAN} = \min_{C} \max_{D} \mathbb{E}\left[\log D(I, M_d) + \log\big(1 - D(C(I, M_i), M_i)\big)\right] \tag{3.13}$$

Joint Loss

In this thesis, the MSE loss function and the GAN loss function are used jointly. The values of the generator loss function and the discriminator loss function are as small as possible, so as to obtain better image inpainting results. The joint loss function is a combined function based on MSE loss L_{mse} and GAN loss, where α refers to the weights of our network. The joint loss function is shown in eq. (3.14).

$$L_{joint}=\min_{C}\ \max_{D}\mathbb{E}\left[\,L_{mse}+\ \log \mathrm{D}(I,M_d)+\ \alpha\log\left(1\ -\ \mathrm{D}(\mathrm{C}(I,M_i\,),M_i\,)\right)\right] \tag{3.14}$$

Although we propose a joint loss function, the MSE loss function needs to be used separately in the training process. Firstly, we add the MSE loss to the training process of the completed network. We then take use of the joint loss function in the discriminator network part for training instead of using the GAN loss function alone. Our training method achieves the effect of improving the training speed and the accuracy of face image inpainting through the joint use of two loss functions.

Algorithm 1 Training process.

Input: Face images from CelebA Dataset.

Output: Face images with mask, repaired images.

```
while iterations $t < T_t(Train)$ do
    Sample 4750 images $I$ from the training data.
    Add random mask $M_i$ to each image $I$.
    If t < 100 then
        The completion network C is updated through the MSE loss function using $(I, M_i)$.
    else
        Generate a mask $M_d$ with random holes for 4750 images $I$.
        Update the discriminators D through the GAN loss function using $(C(I,M_i),M_i)$ and $(I,M_d)$.
        if t > 100 + $T_D$ then
            Update the completion network C through the joint loss using $(I,M_d)$, and $\mathrm{D}(\mathrm{C}(I,M_i),M_i)$.
        end if
    end if
end while
```

3.5 Double-Autoencoder GAN Model

In the double-autoencoder models, we still treat CNN in deep learning as the network foundation and GAN model to construct a complete face image reconstruction network and train our image restoration model through the discriminator and generator in the GAN model, to achieve higher quality.

CNN is the core part of the whole inpainting algorithm, the convolutional layer becomes the key to our whole CNN algorithm. This is the reason why most of the operations in the entire inpainting process are generated in convolutional layers. The generative network consists of double-autoencoders connected. Each autoencoder consists of 12 layers of convolutional operations, 4 layers of dilated convolution, and 2 layers of deconvolution (Gao, Nguyen, & Yan, 2021). The face images generated by double-autoencoders are closer to the "real" images than the images generated using only one autoencoder.

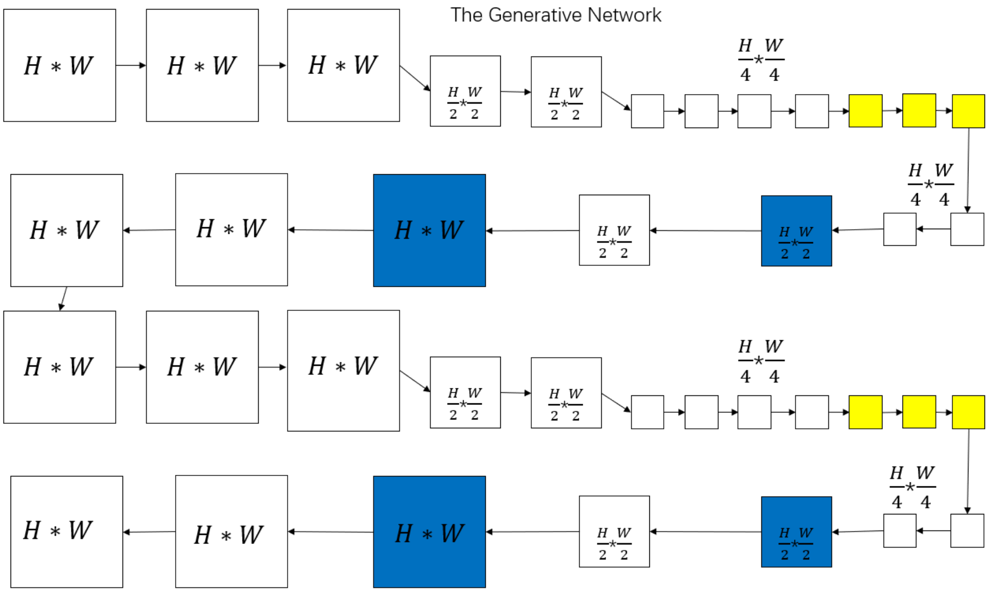
The Generative Network
$H * W$
$H * W$
$H * W$
$\frac{H}{2} * \frac{W}{2}$
$\frac{H}{2} * \frac{W}{2}$
$\frac{H}{4} * \frac{W}{4}$
$\frac{H}{4} * \frac{W}{4}$
$H * W$
$H * W$
$H * W$
$\frac{H}{2} * \frac{W}{2}$
$\frac{H}{2} * \frac{W}{2}$
$H * W$
$H * W$
$H * W$
$\frac{H}{2} * \frac{W}{2}$
$\frac{H}{2} * \frac{W}{2}$
$\frac{H}{4} * \frac{W}{4}$
$\frac{H}{4} * \frac{W}{4}$
$H * W$
$H * W$
$H * W$
$\frac{H}{2} * \frac{W}{2}$
$\frac{H}{2} * \frac{W}{2}$

Figure 3.9: The completion network for double-autoencoder GAN model

The completion network consists of a total of thirty-six layers of convolutional neural networks. There are 34 convolutional layers in the generative network, represented by rectangles with white background color. There have eight yellow rectangles in total, which are dilated convolution, additionally, four additional deconvolution layers are represented by blue squares. The strides of convolutional layer 3, convolutional layer 5, convolutional layer 15 and convolutional layer 17 are both 2×2.

Table 3.4: The completion network

Convolution types		**Parameters**		
	Kernels	***Dilations***	***Strides***	***Outputs***
Convolution layer1	5×5	None	1×1	64
Convolution layer2	3×3	None	1×1	64
Convolution layer3	3×3	None	2×2	128
Convolution layer4	3×3	None	1×1	128
Convolution layer5	3×3	None	2×2	256
Convolution layer6	3×3	None	1×1	256
Convolution layer7	3×3	None	1×1	256
Dilated layer1	3×3	2	1×1	256
Dilated layer2	3×3	4	1×1	256
Dilated layer3	3×3	8	1×1	256
Dilated layer4	3×3	16	1×1	256
Convolution layer8	3×3	None	1×1	256
Convolution layer9	3×3	None	1×1	256
Deconvolution layer1	4×4	None	$\frac{1}{2}\times\frac{1}{2}$	128
Convolution layer10	3×3	None	1×1	128

Convolution types		Parameters		
	Kernels	***Dilations***	***Strides***	***Outputs***
Deconvolution layer2	4×4	None	$\frac{1}{2}\times\frac{1}{2}$	64
Convolution layer11	3×3	None	1×1	32
Convolution layer12	3×3	None	1×1	3
Convolution layer13	5×5	None	1×1	64
Convolution layer14	3×3	None	1×1	64
Convolution layer15	3×3	None	2×2	128
Convolution layer16	3×3	None	1×1	128
Convolution layer17	3×3	None	2×2	256
Convolution layer18	3×3	None	1×1	256
Convolution layer19	3×3	None	1×1	256
Dilated layer5	3×3	2	1×1	256
Dilated layer6	3×3	4	1×1	256
Dilated layer7	3×3	8	1×1	256
Dilated layer8	3×3	16	1×1	256
Convolution layer20	3×3	None	1×1	256
Convolution layer21	3×3	None	1×1	256
Deconvolution layer3	4×4	None	$\frac{1}{2}\times\frac{1}{2}$	128
Convolution layer22	3×3	None	1×1	128
Deconvolution layer4	4×4	None	$\frac{1}{2}\times\frac{1}{2}$	64
Convolution layer23	3×3	None	1×1	32
Convolution layer24	3×3	None	1×1	3

While using double-autoencoder GAN models, the discriminator network is the same as that of the one-autoencoder GAN model. There is still a global discriminator network and a local discriminator network. The global discriminative network consists of five convolutional layers. The local discriminative network consists of four convolutional layers. The global discriminator and the local discriminator are connected through a fully connected layer. The algorithm part is as same as the GAN model composed of one-autoencoders. Both the MSE loss function in the completion network part and the joint loss function in the discriminator network are employed.

3.6 Evaluation Methods

After the image inpainting algorithm runs, the inpainting result is obtained. The quality of the repair results is measured by comparing the difference between the repair results and the undamaged source images. In general, the inpainting algorithm is evaluated from the following two aspects: One is the running time of the algorithm, the other is the evaluation of the repaired image itself.

The running time of the algorithm is an evaluation index for the efficiency of the algorithm. Because the time is a number, it is easy to be quantified and compared. In general, the running time of the algorithm is proportional to the image repair effect, that is, the algorithm with better repair effect takes a relatively long time. As for the evaluation of the repaired image itself, there are few papers covering this aspect, because it is difficult to obtain undamaged images for comparison, so there is no unified evaluation standard, but it can still be evaluated subjectively and objectively.

3.6.1 Subjective evaluation method

Subjective evaluation is to evaluate the repair results through human visual system. The observer can consider the repair of damaged images as a whole, we can clearly see whether there is an error, whether the repaired regions are reasonable, the integrity of the object is to be repaired, whether it can be maintained and so on. However, human vision is subjective, different observers may have distinct evaluation standards for the details of the image brightness, illumination, color, etc. which are not easily observed by human eyes.

Since subjective evaluation is easily influenced by human factors. In this thesis, we design a questionnaire for survey purpose. Three images and the corresponding repaired images of the three methods are randomly selected, the order of the three results of each image is random. There are 30 participants who were invited to conduct a survey of the three groups of results. For scoring, images that people perceive as more natural receive higher scores.

3.6.2 Objective evaluation method

The objective evaluation methods are mainly evaluated by Mean Squared Error (MSE), Peak Signal Noise Rotation (PSNR), and Structural Similarity Index (SSIM).

MSE is an index value applied to calculate the similarity of two images. The smaller the value, the more similar the two images are, which is defined as follows,

$$MSE = \frac{\sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (I_0(i,j) - I(i,j))^2}{M*N} \tag{3.15}$$

MSE represents the mean square error between the pixels of the original image, the restored image. $I_0(i,j)$ shows the value of the pixel at coordinates (i,j) in the image, and $I(i,j)$ indicates the value at this point in the repaired image. M and N are the numbers of rows and columns where the pixel is located, respectively.

PSNR is an evaluation of human perception of reconstruction quality. PSNR is often applied to measure the reconstruction quality of lossy compression codecs, which is also applied to measure the quality of image inpainting. It is simply defined by the mean square error. Generally, a higher PSNR indicates a higher quality of image inpainting which is defined as,

$$PSNR = 10 * \log \left(\frac{\mathrm{MaxI}}{\mathrm{MSE}}\right) \tag{3.16}$$

where MaxI is generally 255, Eq. (3.16) is for the calculation method of the image. If it is a color image, we need to calculate PSNR of the three RGB channels. The result is MSE of the color image. By using PSNR to evaluate the restoration results, there may be a high PSNR, but through visual observation, unreasonable details are found, or the PSNR value is very low, the results observed by the naked eyes are better. However, all the repaired images in the experimental results have their original images, the PSNR value between the two images can be easily calculated, PSNR can be employed as an auxiliary judgment standard.

If PSNR is higher than 40dB, the image quality is particularly good. If the PSNR is between 30dB and 40dB, it usually means the image quality is superiror. If the PSNR is between 20dB and 30dB, the image quality is general. Finally, the image quality is unacceptable if the PSNR is lower than 20dB.

SSIM indicator can better reflect the similarity between the two pictures. SSIM evaluates the similarity of two images in terms of luminance, contrast, and structure. The basic framework of the algorithm is shown as follows,

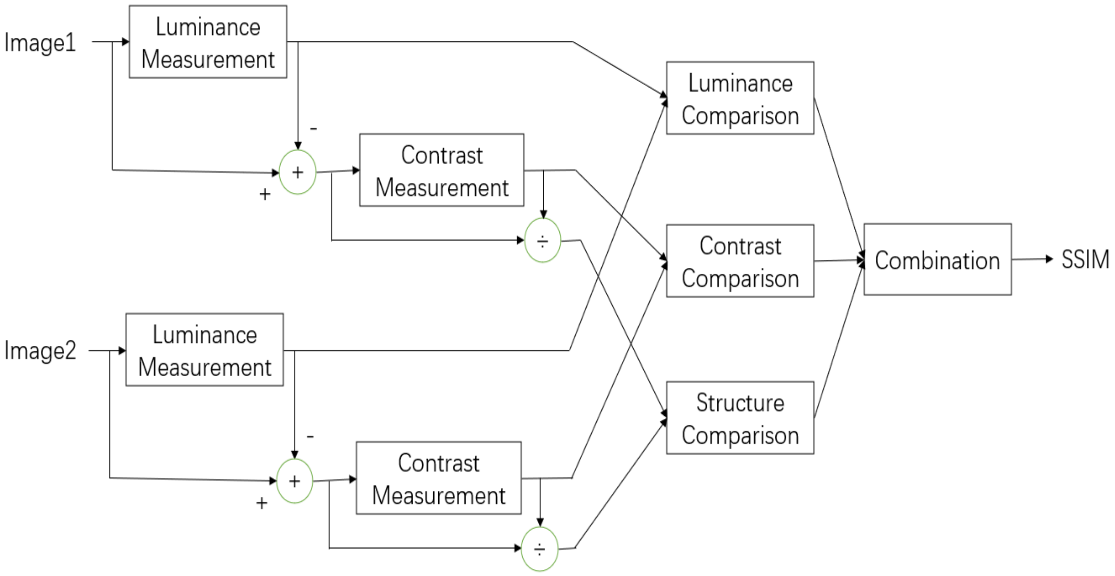

Figure 3.10: The flowchart of SSIM metric

The basic calculation process of SSIM metric is to firstly calculate the luminance measurement and make a comparison to get the first similarity-related evaluation. After subtracted the effect of luminance, the contrast measurement is calculated, a second evaluation is obtained by comparison. Finally, we take use of the results obtained in the previous step to remove the influence of contrast and perform structure comparison. The three obtained results are mixed to obtain the final evaluation result.

The range of the indicator is (0,1). If SSIM = 0, it means that the two images are completely dissimilar. If SSIM = 1, the two images are very similar. That is, the closer the value is to 1, the more similar the two pictures are, eq. (3.17) is shown for calculating this indicator as follows,

$$SSIM(x,y) = \frac{(2\mu x\mu y+c1)(2\sigma xy+c2)}{(\mu x2+\mu y2+c1)(\sigma x2+\sigma y2+c2)} \tag{3.17}$$

where μ_x represents the mean value in the horizontal direction in the *N* x *N* image, σ_x represents the variance in the horizontal direction, c_1 and c_2 show the average pixel intensity of the two test images, respectively, σ_{xy} reflects the covariance in the horizontal and vertical directions.

Table 3.5 Parameter ranges

Parameters	**Ranges**
MSE	(0,1)
PSNR	$(0,\infty)$
SSIM	(0,1)

From MSE and PSNR formulas, PSNR is calculated based on the MSE. Therefore, we will take advantage of PSNR and SSIM for the purpose of comparison in the analysis of the results, the graph of experimental results will be judged by thirty participants.

Chapter 4
Results

The main content of this chapter is the experimental results and comparisons of face image inpainting. Finally, in this chapter, we will discuss the limitations of this project.

4.1 Data Collection and Experimental Environment

In Section III, we propose a new GAN model. The MSE loss and GAN loss are employed in the loss function part. Firstly, we only take use of the MSE loss in the first hundred iterations and add the GAN loss for training. For the convenience of comparison, we train the network for five hundred epochs.

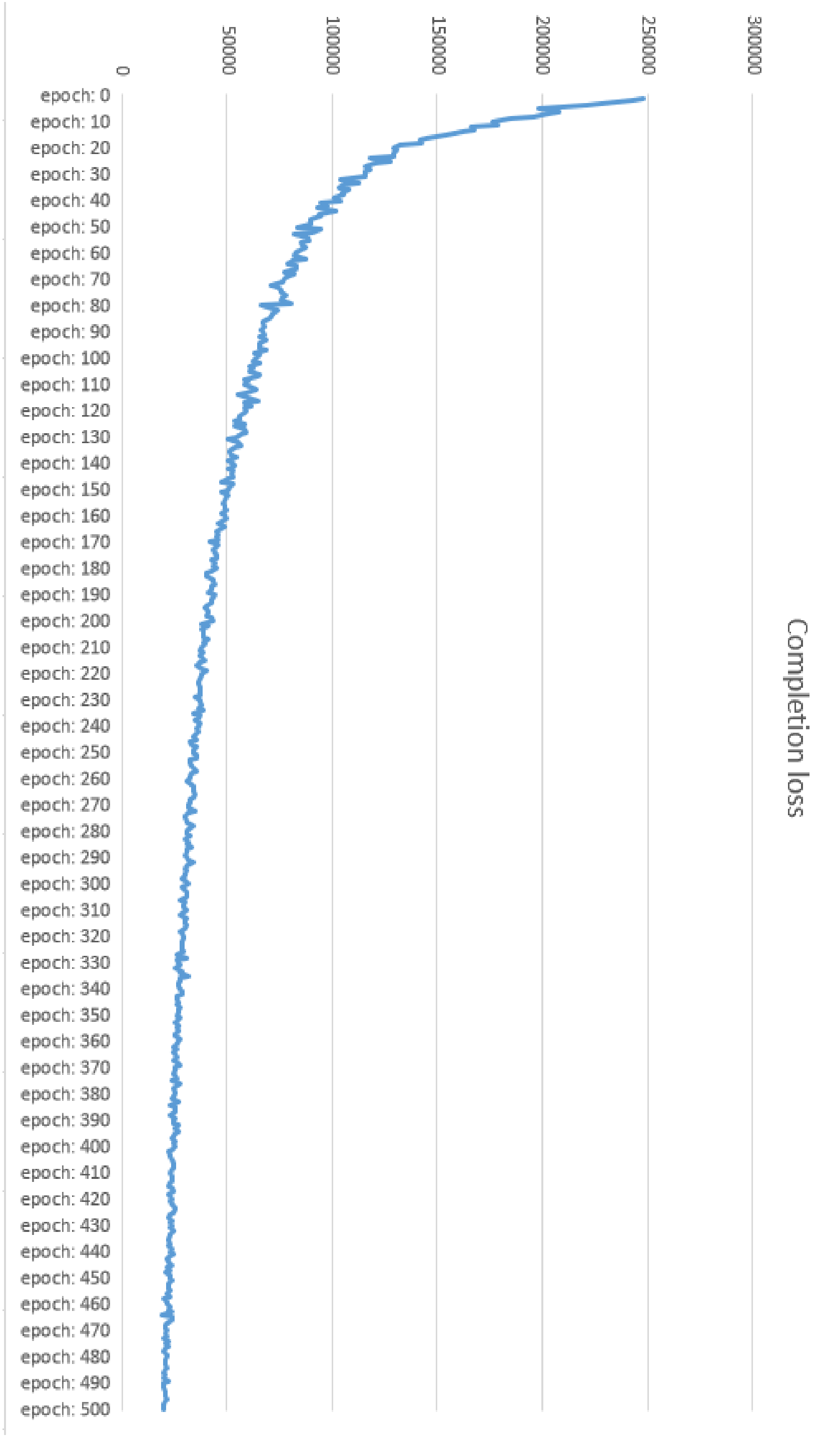

Figure 4.1: The completion loss

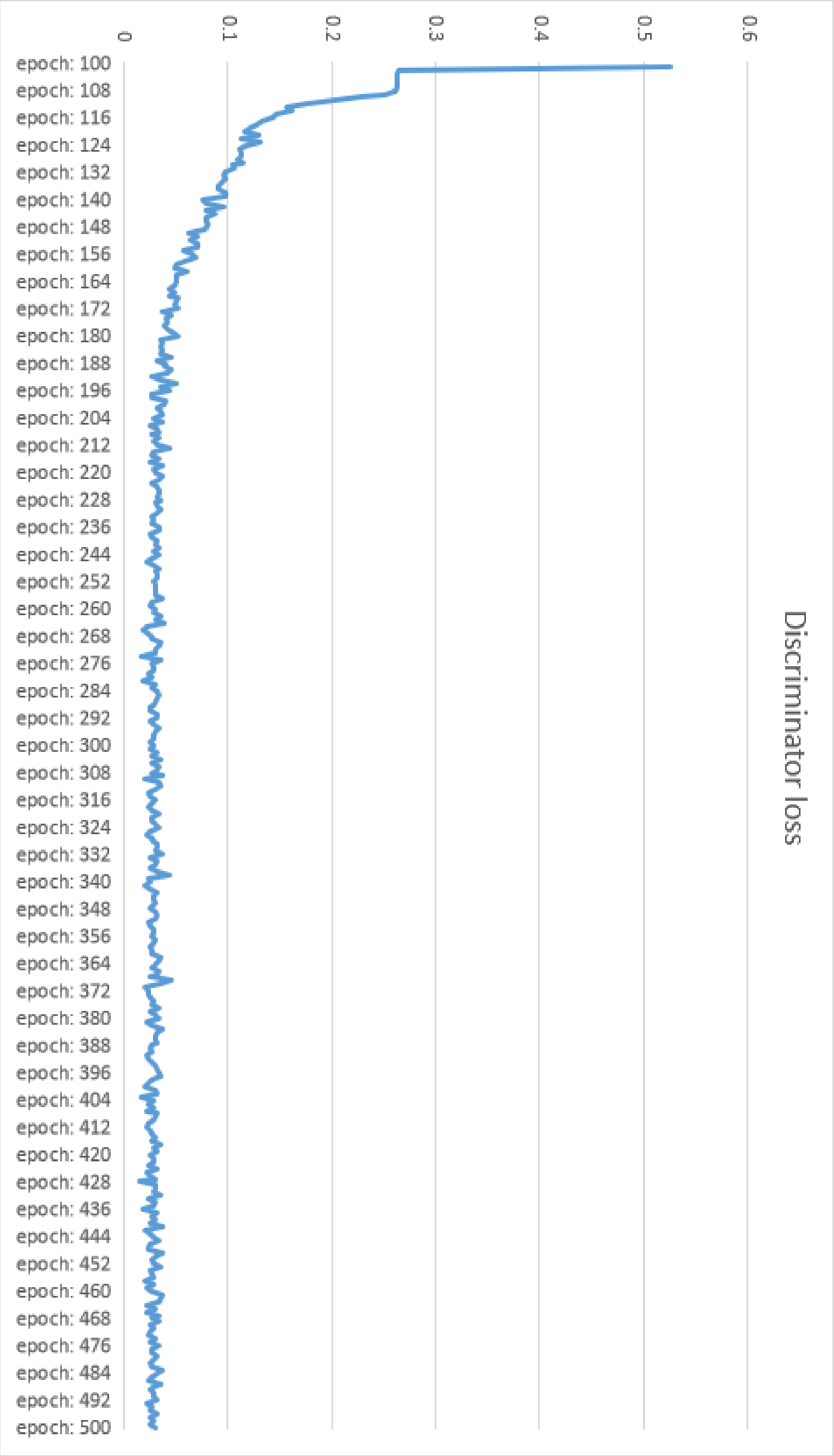

Figure 4.2: The discriminator loss

The inpainting results are shown in Figure 4.5. We see that the inpainting is completed by using autoencoder if we randomly add a mask to the image. The inpainted image is very close to the result of the full image. With our naked eyes, this is undoubtedly a very successful restoration. However, we still need data to support the conclusions we see, hereinafter we choose MSE, PSNR and SSIM as our data support. Since PSNR is calculated based on MSE. So, we are use of PSNR and SSIM as evaluation metrics in the testing phase. PSNR is the most popular objective metric for evaluating image quality. In general, the larger the PSNR value, the higher the quality of the image inpainting. SSIM is a measure of how similar two images are (Sara, Akter, & Uddin, 2019). The result is between 0 and 1.00, the closer the result is to 1.00, the higher the quality of the image inpainting.

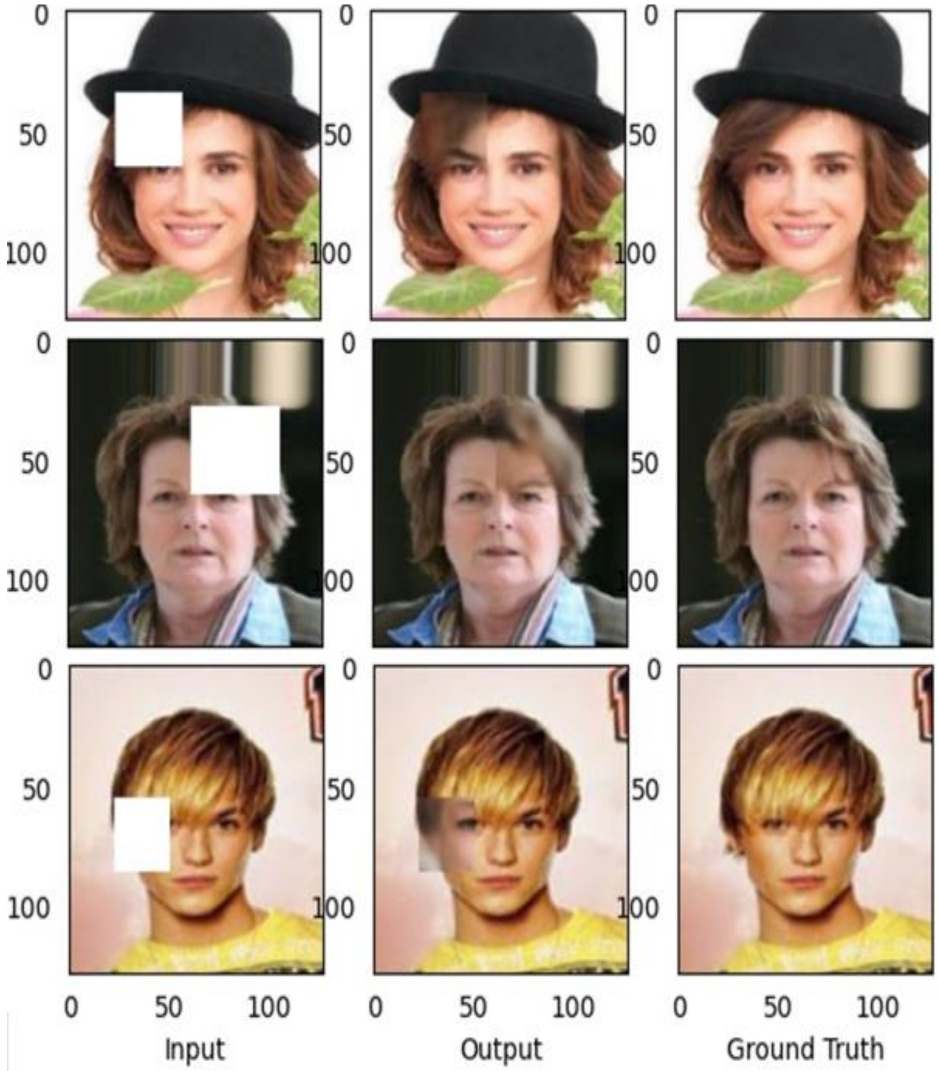

Figure 4.3: New model test results

The completion network is trained for 100 iterations, the discriminator is then trained for T_D= 400 iterations, finally, both are jointly trained to reach a total of $T_t(Train)$ = 500 iterations. On a single machine with only one 3060 GPU, the entire training process takes around 4 days.

We are use of images from faces, not in the training data to evaluate our model and compare with the existing methods to demonstrate the performance of our approach. Our model is trained based on the CelebA dataset.

In this thesis, we evaluate the effect of repair on arbitrary regions. We compare our results with those results (Pathak et al. 2016) (Iizuka et al. 2017). For comparisons, we retrained the two models. During training, we are use of the same epoch and add an arbitrary mask. All input images are fixed to 128×128 for our evaluations. Through image preprocessing and model training, the repaired image is obtained. In addition, PSNR and SSIM were selected for result analysis in the evaluation.

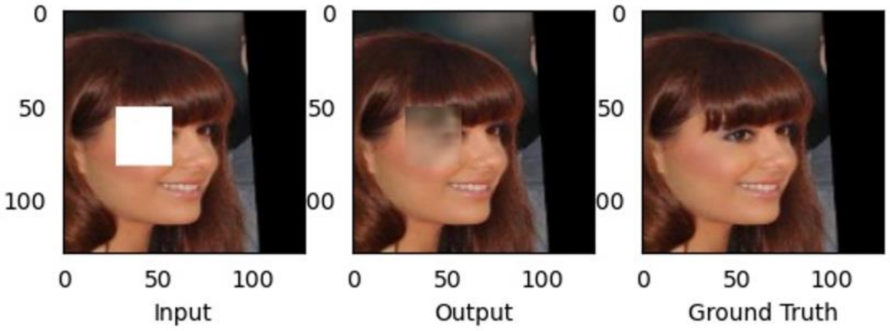

Figure 4.4: The example of inpainting result (Pathak et al.2016)

In Figure 4.4, the experimental results (Pathak et al.2016) are relatively blurry to human eyes, the experiment pays much attention on the influence of global information on the image.

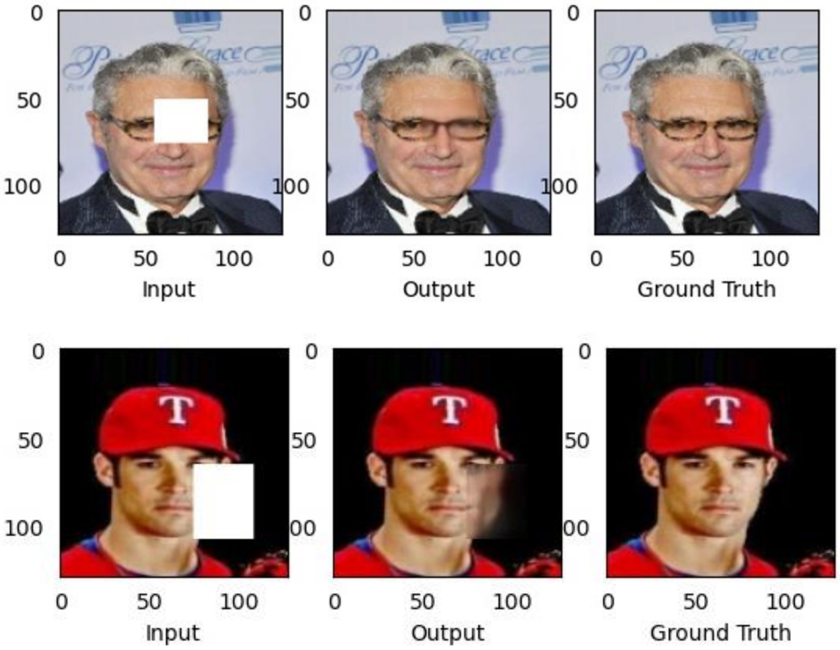

Figure 4.5: The results of local discriminator

The training results are then compared with the full methods for evaluating the impact of global and local discriminators. We show the results of local discriminator in Figure 4.5. The local discriminator only takes use of the nearby pixels in the process of inpainting the image. From Figure 4.5, we see that a part of the face has been repaired, but it is more than the one on the right, resulting in asymmetry on both sides.

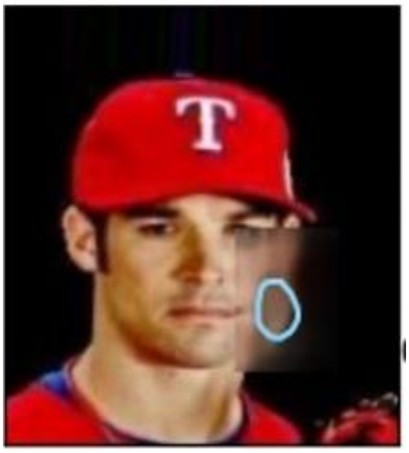

Figure 4.6: Failure of local discriminator in image inpainting

The part circled in blue is the part with poor repair effect. We show the processing results of the global discriminator in Figure 4.7. The image is more symmetrical in the result of the global discriminator repair. The results of the repair are globally good, and the left and right eyes after repair are relatively similar.

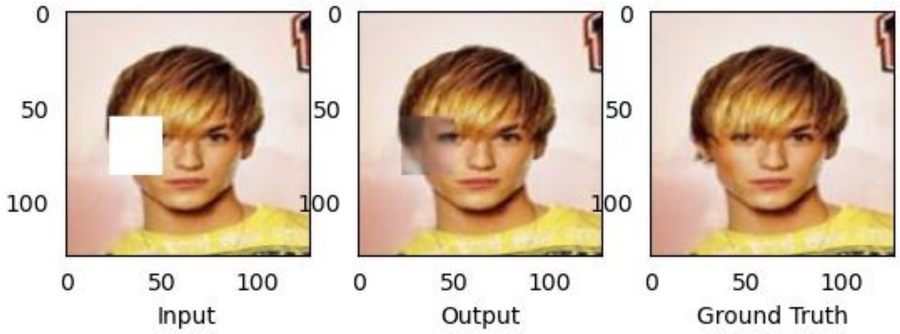

Figure 4.7: The global discriminator results

Table 4.1 Local and global comparison

Parameters	Local	Global
PSNR	13.00	19.82
SSIM	0.88	0.90

By using both global and local discriminators, we are able to achieve locally and globally consistent inpainting results.

The masked locations and the corresponding results of face image were inpainting by using the trained model of Iizuka's net are shown in Fig 4.8. The first column is the input images with binary marks, the second column is the output result, and the third column is the ground truth.

Four mask positions and corresponding results of face image inpainting by using one-autoencoder GAN net are shown in Fig. 4.9. The first column is the input image with binary labels, the second column is the output result, and the third column is the ground truth.

It is worth mentioning that the input image size in the dataset is small. Compared to Iizuka's network, we added one more convolutional network to the completion network and reduce the convolutional operations per discriminator network by using one layer.

Therefore, the time required for training is cut down. It only takes five hundred epochs to get a well-trained network from training images. The four sets of images in Figure 4 show the output of Iizuka's network's well-trained algorithm for face image inpainting: Input image, output image, and image sequence. As a comparison, using the same dataset for five hundred epochs, the results obtained by the one-Autoencoder GAN model network are shown in Figure 4.9.

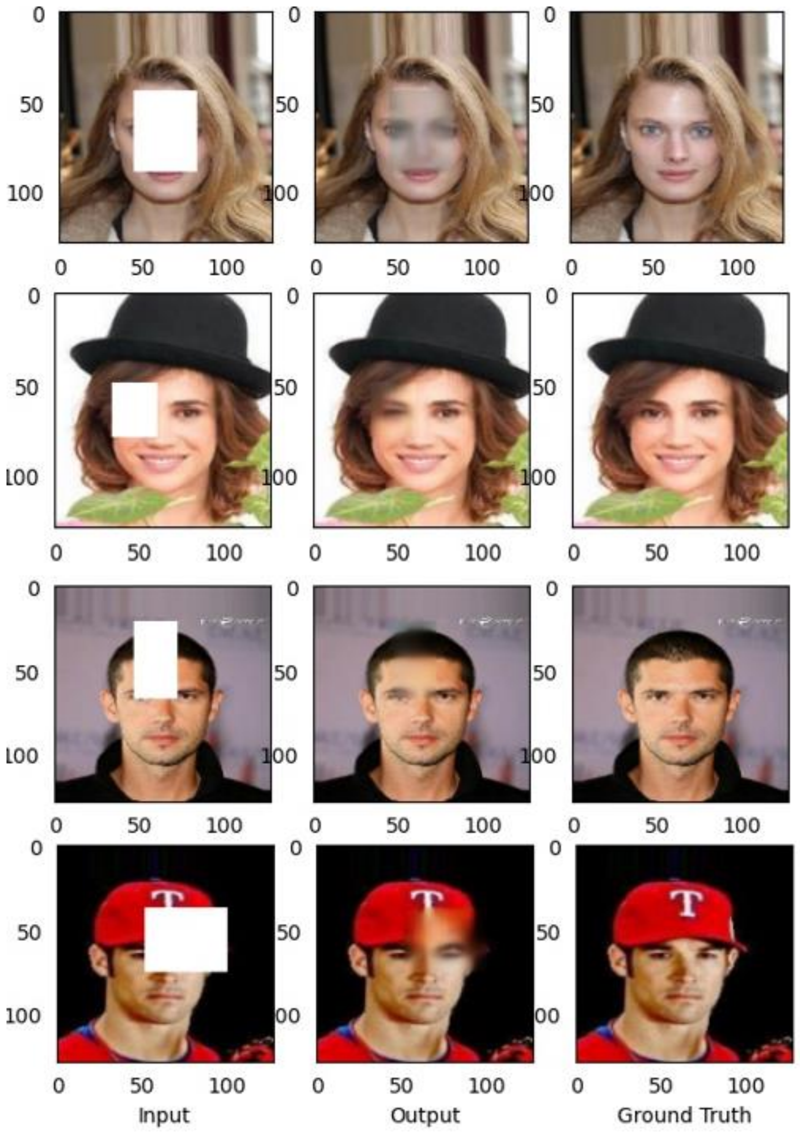

Figure 4.8: The results from Iizuka’s net

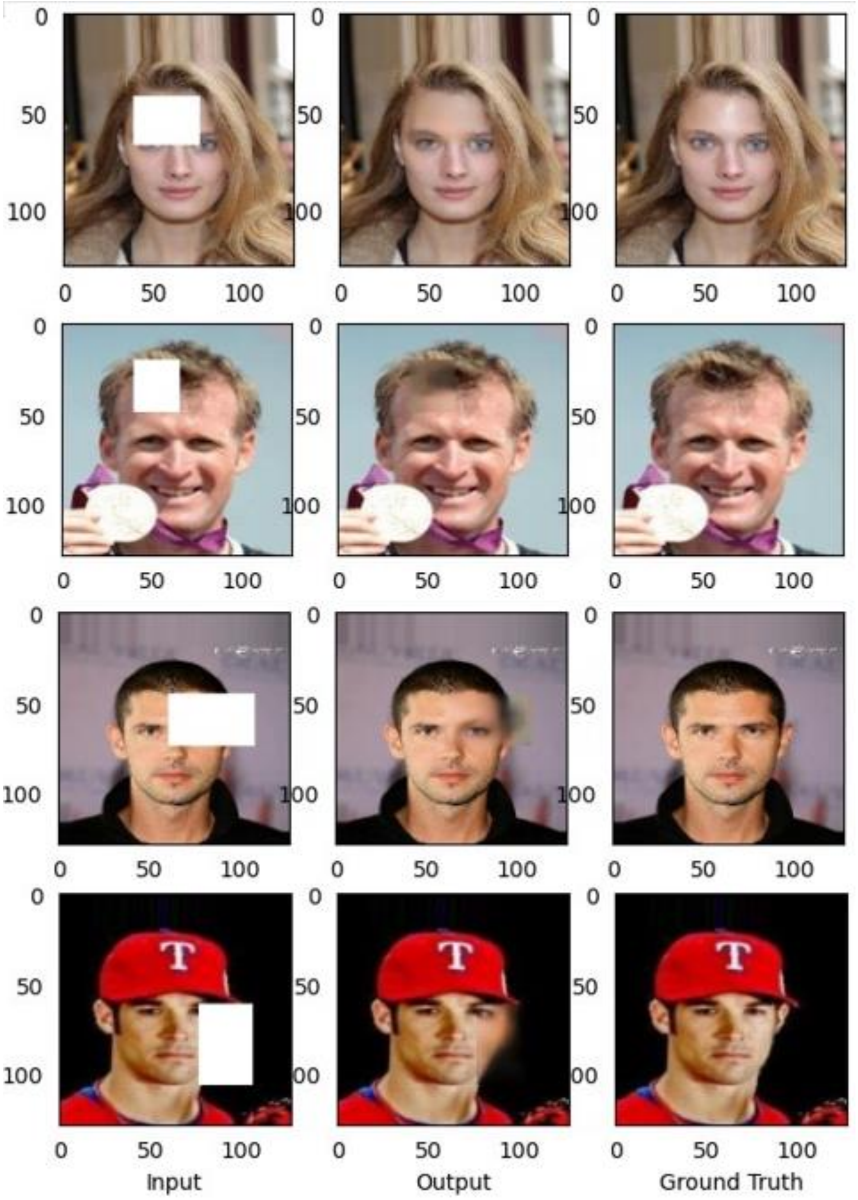

Figure 4.9: One-autoencoder GAN model results

Table 4.2 Results comparison

Iizuka's results		
Names	***PSNRs***	***SSIMs***
Result1	17.76	0.80
Result2	19.54	0.80
Result3	16.64	0.83
Result4	14.39	0.81
Average	**17.09**	**0.84**

Our results (One-Autoencoder GAN)		
Names	***PSNRs***	***SSIMs***
Result1	58.13	0.98
Result2	18.40	0.77
Result3	31.44	0.95
Result4	19.47	0.86
Average	**31.86**	**0.89**

We see that our proposed network obtains an average of 31.86 for PSNR and 0.89 for SSIM in the test results obtained by our proposed network. After training Iizuka's network for five hundred epochs, the average PSNR in the test results is 17.09, the average SSIM is 0.84. By comparing the results obtained with other trained models, we found that the image quality of our modified model has improved.

Four mask positions and corresponding results of face image inpainting by using double-autoencoder GAN net. The experimental results of the GAN model network of double-autoencoder are shown in Figure 4.10. The resulting images have an average PSNR of 36.74. The average result of another parameter SSIM is 0.91. Compared with the other two network models, there is a significant improvement.

Another major motivation for image inpainting is to remove unwanted parts of an image. We show an example in Figure 4.11. The woman's sunglasses are removed, and the face part can be repaired. We see that the eye part has been successfully recovered in the result. Our results obviously are correct.

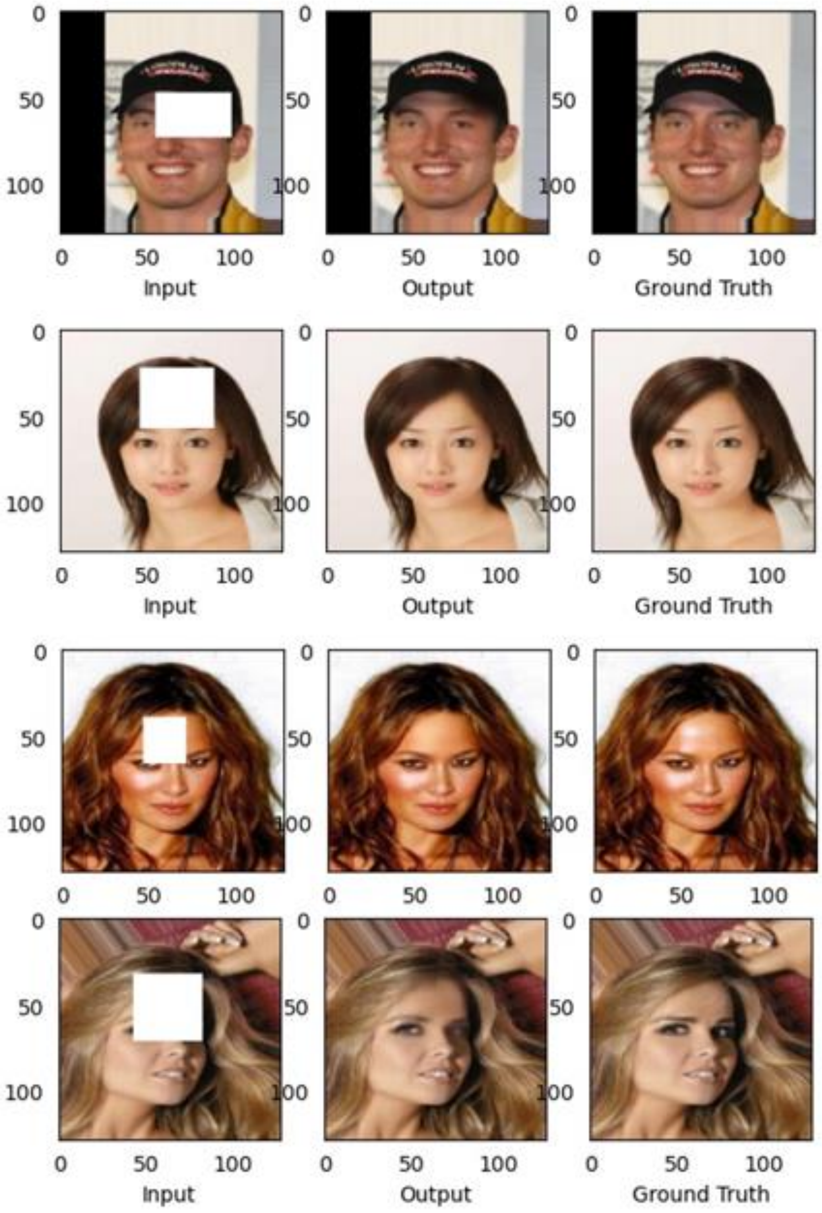

Figure 4.10: Double-Autoencoder GAN model Results

Table 4.3 Results comparison (Iizuka's net, one-Autoencoder GAN model and double-Autoencoder GAN model)

Model names Parameters	PSNR	SSIM
Iizuka et al. model	17.09	0.84
Our one-Autoencoder GAN	31.86	0.89
Our double-Autoencoder GAN	36.74	0.91

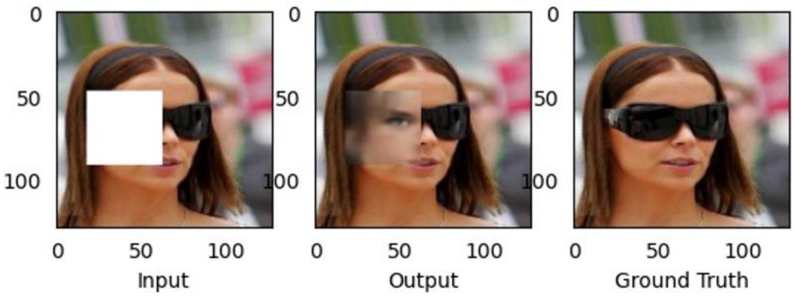

Figure 4.11: Remove sunglasses

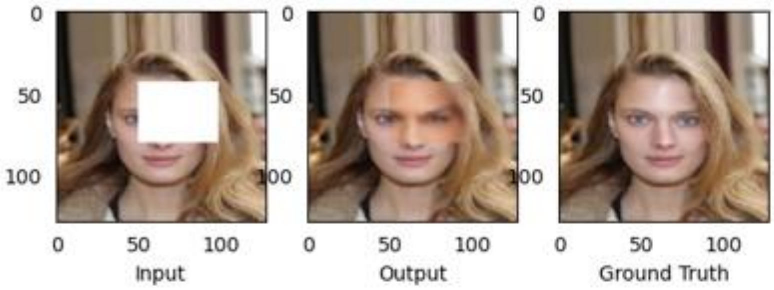

Figure 4.12: The failed image inpainting results

There are drawbacks in our proposed algorithm. In Figure 4.12, the female's restored eye color is different from the original color. The woman's original eye color was blue, and the finished color was black. This is a failed restoration image.

4.2 Limitations of the Research Work

Our proposed algorithm has been successfully applied in face image inpainting. But there are still a number of limitations that need to be improved. The restrictions include:

Only a small number of face images are included in the dataset, which makes our model less effective at inpainting some faces. Since we only used the CelebA dataset for training, the model currently only works well on this dataset.

Our double-autoencoder GAN model adds more convolutional neural networks to the completion network, which runs slower than the pne-autoencoder GAN model.

The image restoration takes use of 5,000 images in the CelebA dataset. Since most of the face images in the CelebA dataset do not have masks, we cannot get excellent image inpainting results with mask removal in the test. Face inpainting without masks also needs to add more face images with masks. Considering the particularity of each face, we will put more face images into the training dataset in the future.

Chapter 5
Analysis and Discussions

The focus of this chapter is on the analysis and comparison of experimental results. The results of three different GAN models are compared.

5.1 Analysis

The purpose of this thesis is to improve the accuracy of face image inpainting. We introduce face image inpainting with Python. At present, the face image inpainting datasets are CelebA and CelebA-HQ. During the training process, we selected the CelebA dataset because this dataset has pretty rich data and features. The three evaluation methods for image inpainting are MSE (Palubinskas, 2017), PSNR and SSIM. Since PSNR is calculated by MSE, we only show the results of PSNR and SSIM in the analysis process. The image restoration takes advantage of the known area information in the image to fill the damaged area, the image restoration is as similar as possible to the original image. Therefore, the quality of the image restoration results can also be subjectively judged by our human eyes.

5.1.1 PSNR

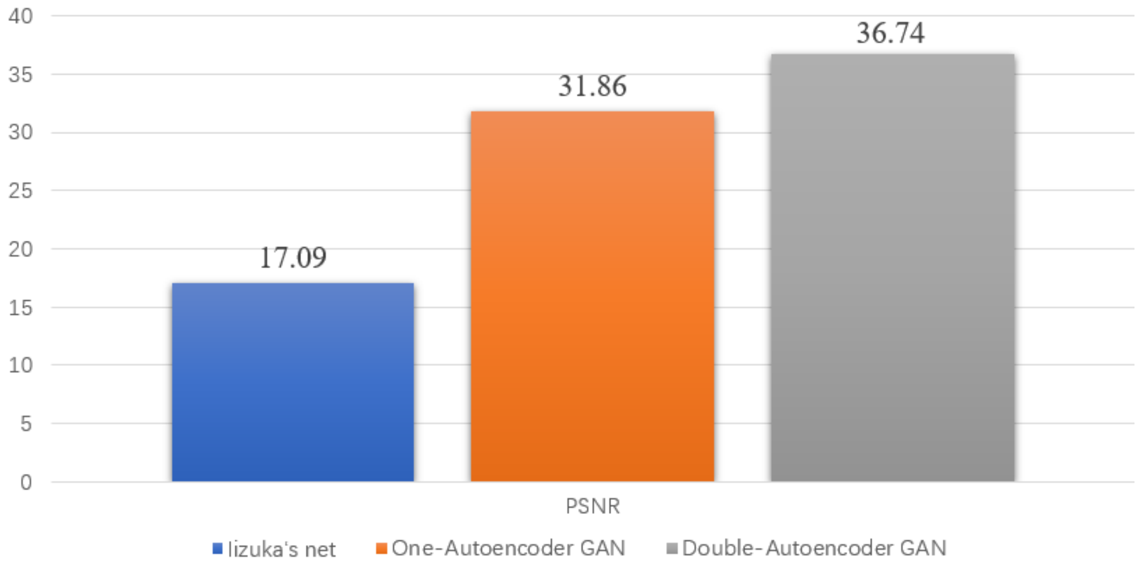

Figure 5.1: PSNR of the three models

After trained Iizuka's network for 500 epochs, the average PSNR values in the test results is 17.09. After trained with the same epoch, the average PSNR in the test results obtained by using our proposed one-autoencoder GAN model is 31.86. Another model we proposed, double-autoencoder GAN, received an average PSNR 36.74. By comparing the results obtained by the three trained models, we see that the image quality of our modified model has been improved.

5.1.2 SSIM

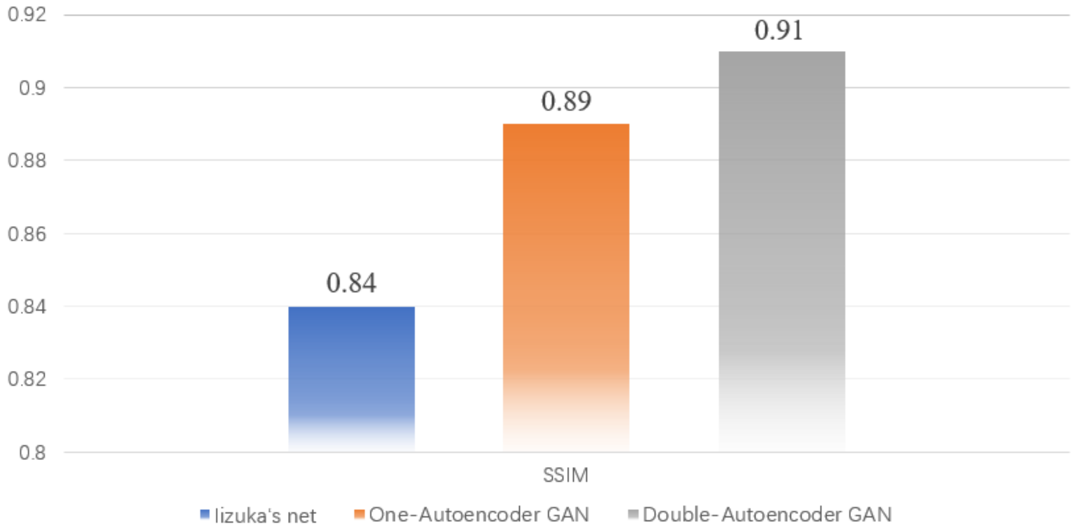

Figure 5.2: SSIM of the three models

The average value of SSIM in the test results of Iizuka's network is 0.84. After trained the same epoch, the average SSIM in the test results obtained by our proposed one-Autoencoder GAN model is 0.89. The average SSIM obtained by using double-autoencoder GAN model is 0.91. By comparing SSIM, we see that the repaired results of images by using the double-autoencoder GAN model is the best.

5.1.3 Subjective Evaluations of Human Eyes

Result evaluations using human eyes are fully subjective (Xu, Chen, Zhang, & Wu, 2011), the accuracy is not as accurate as PSNR and SSIM. We found 30 different users, given them the experimental results of the proposed two models, and required their feedback. Based on their feedback, we made a boxplot. Figure 5.3 shows the results of a survey evaluating the naturalness of images.

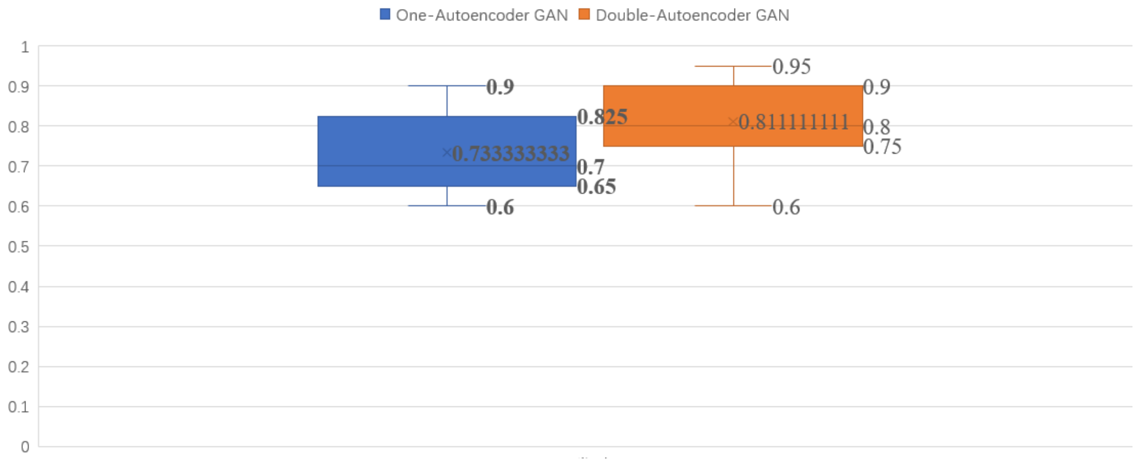

Figure 5.3: The boxplot of user ratings

The results in Figure 5.3 show that 30 users felt that the inpainted images obtained by the Double-Autoencoder GAN model were more natural.

5.2 Discussions

In the experiments, we compared the three models. The proposed model has received average PSNR 17.09 and SSIM 0.84. The network with only one autoencoder in the generative network has an average PSNR 31.86 and SSIM 0.89. The test results for a network with two autoencoders in the generative network have an average PSNR 36.74 and SSIM 0.91. We see that the model using two autoencoders has a higher PSNR compared to the model using one autoencoder, our model is about 5% higher based on this metric.

The same is true for another metric, SSIM is 0.02 higher than the model by using one autoencoder. By comparisons, we find that the double-autoencoder GAN model has better inpainting performance. Furthermore, the SSIM of the double-autoencoder GAN model is as high as 0.91, which is very close to 1.00. The PSNR is also high, which not only proves that our results are very good, but also further indicates the reliability of these two metrics. We also obtained a more natural restoration result obtained by the Double-Autoencoder GAN model through the evaluation of the naturalness of the image by 30 participants.

Overall, our adopted deep learning model can accurately inpaint missing parts of face images. The more iterations, the better the visual effect. Although the double-autoencoder GAN model takes a long time in the training process, it is still a valid model. In addition, the results of the model's evaluation metrics SSIM and PSNR are also excellent, it looks natural and realistic for the repair effect.

Chapter 6
Conclusion and Future Work

In this chapter, we will summarize the subject and method of this project and propose a new research direction according to the results and insufficiency of our experiments, preparing for the future work.

6.1 Conclusion

Due to Covid-19, people are wearing masks when traveling. Unlocking a phone while wearing a mask is a challenge. With the rapid development of deep learning in the past few years, more deep learning methods are employed in image inpainting. Face images inpainting with masks can help solve the problem of unlocking mobile phones.

In this thesis, autoencoder-based and GAN-based face image inpainting methods are investigated. We proposed a new deep learning model. The structure of the overall model is similar to the GAN model. However, two autoencoder networks were employed in the completion network part. We took use of a local discriminator and a global discriminator in the discriminator network. We also modified the loss function in the training process, first using the MSE loss function in the complete network part, and then using a joint loss function consisting of the MSE loss function and the GAN loss function in the discriminator network. Unfortunately, the training speed of the model slows down due to the addition of convolutional layers that complete the network part of the model. Nevertheless, the improved model and loss function improve the quality of image inpainting. The evaluation parameter PSNR of the inpainting results is improved to 36.74dB, SSIM is improved to 0.91, and the inpainting results look more natural.

6.2 Future Work

In the future, based on current results, we will employ other testing methods to further demonstrate the superiority of our algorithms. In addition, we will continue to optimize the algorithms to achieve better results. We will increase the speed of repairs without degrading the visuals. Our current limitation is that the double-autoencoder GAN model is slow to train, and we will try to improve the model to increase the speed of face image inpainting without reducing image quality.

Furthermore, our limitation is that current public datasets of faces are limited. Moreover, the current face datasets are all datasets without masks. We will collect more face images with masks to make a dataset. We will test more datasets and check the results of our proposed method. We will impaint face images with a large number of missing regions to improve the generality of these models.

References

Al-Sarayreha, M. (2020) Hyperspectral Imaging and Deep Learning for Food Safety. PhD Thesis. Auckland University of Technology, New Zealand.

An, N., Yan, W. (2021) Multitarget tracking using Siamese neural networks. *ACM Transactions on Multimedia Computing, Communications and Applications.*

An, N. (2020) *Anomalies Detection and Tracking Using Siamese Neural Networks.* Master's Thesis. Auckland University of Technology, New Zealand.

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. International Conference on Engineering and Technology, pp. 1-6.

Atrey, P., Yan, W., Chang, E., Kankanhalli, M. (2004) A hierarchical signature scheme for robust video authentication using secret sharing. *International Multimedia Modelling Conference*, 330-337.

Atrey, P., Yan, W., Kankanhalli, M. (2007) A scalable signature scheme for video authentication. Multimedia Tools and Applications 34 (1), 107-135.

Bansal, M., Yan, W., Kankanhalli, M. (2003) Dynamic watermarking of images. International Conference on Information, Communications and Signal Processing.

Bengio, Y., Yao, L., Alain, G., & Vincent, P. (2013). Generalized denoising auto-encoders as generative models. Advances in neural information processing systems, pp. 899-907.

Bertalmio, M., Sapiro, G., Caselles, V., & Ballester, C. (2000). Image inpainting. Annual conference on Computer graphics and interactive techniques, pp. 417-424.

Boureau, Y. L., Le Roux, N., Bach, F., Ponce, J., & LeCun, Y. (2011). Ask the locals: Multi-way local pooling for image recognition. International Conference on Computer Vision, pp. 2651-2658.

Boureau, Y. L., Ponce, J., & LeCun, Y. (2010). A theoretical analysis of feature pooling in visual recognition. International Conference on Machine Learning, pp. 111-118.

Cao, X. (2022) Pose Estimation of Swimmers from Digital Images Using Deep Learning. Master's Thesis, Auckland University of Technology.

Chambers, J., Yan, W., Garhwal, A., Kankanhalli, M. (2014) Currency security and forensics: A survey. Multimedia Tools and Applications, 74(11), 4013-4043.

Chan, T. F., & Shen, J. (2001). Nontexture inpainting by curvature-driven diffusions. Journal of visual communication and image representation, pp. 436-449.

Chan, T. F., & Shen, J. (2000). Mathematical models for local deterministic inpaintings. UCLA CAM TR.

Chen, S. S., Donoho, D. L., & Saunders, M. A. (2001). Atomic decomposition by basis pursuit. SIAM review, pp. 129-159.

Chen, Y., Zhang, H., Liu, L., Chen, X., Zhang, Q., Yang, K., & Xie, J. (2021). Research on image inpainting algorithm of improved GAN based on two-discriminations networks. Applied Intelligence, pp. 3460-3474.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. IEEE Signal Processing, pp. 53-65.

Criminisi, A., Perez, P., & Toyama, K. (2003). Object removal by exemplar-based inpainting. In IEEE Conference on Computer Vision and Pattern Recognition, pp. II-II.

Cui, W., Yan, W. (2016) A scheme for face recognition in complex environments. International Journal of Digital Crime and Forensics (IJDCF) 8 (1), 26-36.

Cui, W. (2015) A Scheme of Human Face Recognition in Complex Environments. Master's Thesis, Auckland University of Technology.

Dargan, S., Kumar, M., Ayyagari, M. R., & Kumar, G. (2020). A survey of deep learning and its applications: A new paradigm to machine learning. Archives of Computational Methods in Engineering, pp. 1071-1092.

Denton, E. L., Chintala, S., & Fergus, R. (2015). Deep generative image models using a Laplacian pyramid of adversarial networks. Advances in Neural Information Processing systems.

Ding, W., Yan, W., Qi, D. (1999) Digital watermark image embedding based on U-system. *International Conference on Computer Aided Design and Computer Graphics*, 893-899.

Ding, W., Yan, W., Qi, D. (1999) Digital image scrambling based on Gray code. *International Conference on CAD/CG* 3, 900-904.

Ding, W., Yan, W. (1999) Digital watermark image based on discrete cosine transform. *Journal of North China University of Technology China*.

Ding, W., Yan, W., Qi, D. (2000) Digital image information hiding technology and its application based on scrambling and amalgamation. *Journal of Image and Graphics* 5 (8), 644-649.

Ding, W., Yan, W., Qi, D. (2000) Digital image information hiding technology and its application based on scrambling and amalgamation. *Chinese Journal of*

Computers 10, 644-649.

Ding, W., Yan, W., Qi, D. (2000) Digital image scrambling and digital watermarking technology based on Conway's game. *Journal of North China University of Technology* 12 (1), 1-5.

Ding, W., Yan, W., Qi, D. (2000) Cox's and Pitas's schemes for digital image watermarking. *Journal of Northern China University of Technology* 12 (3), 1-12.

Ding, W., Yan, W., Qi, D. (2000) Digital image scrambling and digital watermarking technology based on Conway's Game. *International Conference on Image Processing.*

Ding, W., Yan, W., Qi, D. (2001) Digital image scrambling. *Progress in Natural Science* 11 (6), 454-460.

Ding, W., Yan, W., Qi, D. (2001) Digital image watermarking based on U-system. *Journal of Image and Graphics* 6 (6), 552-557.

Ding, W., Yan, W., Qi, D. (2001) Digital image scrambling technology based on Arnold transformation. *Journal of Computer Aided Design & Computer Graphics* 13 (4), 338 -342.

Ding, W., Yan, W., Qi, D. (2002) Digital image watermarking based on discrete wavelet transform. *Journal of Computer Science and Technology* 17 (2), 129-139.

Feng, H., Ling, H., Zou, F., Yan, W., Lu, Z. (2010) Optimal collusion attack for digital fingerprinting. *ACM International Conference on Multimedia*, 767-770.

Feng, H., Ling, H., Zou, F., Yan, W., Lu, Z. (2012) A collusion attack optimization strategy for digital fingerprinting. *ACM Transactions on Multimedia Computing, Communications, and Applications*.

Feng, H., Ling, H., Zou, F., Yan, W., Sarem, M., Lu, Z. (2013) A collusion attack optimization framework toward spread-spectrum fingerprinting. Applied Soft Computing 13 (8), 3482-3493.

Finn, C., Christiano, P., Abbeel, P., & Levine, S. (2016). A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. ArXiv preprint.

Frid-Adar, M. D., Goldberger, J., & Greenspan, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing, pp. 321-331.

Fu, Y., Nguyen, M., Yan, W. (2022) Grading methods for fruit freshness based on deep learning. *Springer Nature Computer Science.*

Fu, Y. (2020) *Fruit Freshness Grading Using Deep Learning.* Master's Thesis. Auckland University of Technology, New Zealand.

Gao, X., Nguyen, M., & Yan, W. Q. (2021). Face image inpainting based on generative adversarial network. International Conference on Image and Vision Computing New Zealand, pp. 1-6.

Gao, X., Nguyen, M., Yan, W. (2022) A face image inpainting method based on autoencoder and adversarial generative networks. Pacific-Rim Symposium on Image and Video Technology.

Garhwal, A., Yan, W. (2015) Evaluations of image degradation from multiple scan-print. *International Journal of Digital Crime and Forensics* (IJDCF) 7 (4), 55-65.

Garhwal, A., Yan, W., Narayanan, A. (2017) Image phylogeny for simulating multiple print-scan. *International Conference on Image and Vision Computing New Zealand* (IVCNZ).

Garhwal, Abhimanyu Singh (2018) *Bioinformatics-Inspired Analysis for Watermarked Images with Multiple Print and Scan.* PhD Thesis, Auckland University of technology, New Zealand.

Garhwal, A., Yan, W. (2018) BIIIA: A bioinformatics-inspired image identification approach, *Multimedia Tools and Applications*.

Gowdra, N. (2021) *Entropy-Based Optimization Strategies for Convolutional Neural Networks.* PhD Thesis, Auckland University of Technology, New Zealand.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, pp. 2672-2680.

Gou, C., Wu, Y., Wang, K., Wang, K., Wang, F. Y., & Ji, Q. (2017). A joint cascaded framework for simultaneous eye detection and eye state estimation. Pattern Recognition, pp. 23-31.

Gu, D., Nguyen, M., Yan, W. (2016) Cross models for twin recognition. *International Journal of Digital Crime and Forensics* 8 (4), 26-36.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., & Chen, T. (2018). Recent advances in convolutional neural networks. Pattern Recognition, pp. 354-377.

Gu, Q., Yang, J., Kong, L., Yan, W., Klette, R. (2017) Embedded and real-time vehicle detection system for challenging on-road scenes. *Optical Engineering*, 56 (6), 063102.

Gu, Q., Yang, J., Yan, W., Klette, R. (2017) Integrated multi-scale event verification in an augmented foreground motion space. Pacific-Rim Symposium on Image and

Video Technology (pp.488-500)

Gu, Q., Yang, J., Yan, W., Li, Y., Klette, R. (2017) Local Fast R-CNN flow for object-centric event recognition in complex traffic scenes. Pacific-Rim Symposium on Image and Video Technology (pp.439-452)

Han, C., & Wang, J. (2021). Face image inpainting with evolutionary generators. IEEE Signal Processing Letters, pp. 190-193.

He, C., Hu, C., Zhang, W., & Shi, B. (2014). A fast adaptive parameter estimation for total variation image restoration. IEEE Transactions on Image Processing, pp. 4954-4967.

Ho, J., & Ermon, S. (2016). Generative adversarial imitation learning. Advances in Neural Information Processing Systems, pp. 4565–4573.

Hu, W., & Tan, Y. (2017). Generating adversarial malware examples for black-box attacks based on GAN. arXiv:1702.05983v1

Iizuka, S., Simo-Serra, E., & Ishikawa, H. (2017). Globally and locally consistent image completion. ACM Transactions on Graphics, pp. 1-14.

Jam, J., Kendrick, C., Walker, K., Drouard, V., Hsu, J. G., & Yap, M. H. (2021). A comprehensive review of past and present image inpainting methods. Computer Vision and Image Understanding, pp. 103-147.

Jia, N., Zheng, C., & Sun, W. (2019). A model of emotional speech generation based on conditional generative adversarial networks. In International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), pp. 106-109.

Jiao, Y., Weir, J., Yan, W. (2011) Flame detection in surveillance. Journal of Multimedia 6 (1).

Kieran, D., Wang, Y., Fennell, D., Quinn-O'Brien, J., Yan, W., Crookes, D. (2012) Whole slide imaging in digital pathology: The impact of image compression. *International Congress on Virtual Microscopy.*

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, pp. 1-9.

Laadjel, M., Bouridane, A., Kurugollu, F., Nibouche, O., Yan, W. (2010) Partial palmprint matching using invariant local minutiae descriptors. *Transactions on Data Hiding and Multimedia Security V.*

Laadjel, M., Kurugollu, F., Bouridane, A., Yan, W. (2019) Palmprint recognition based on subspace analysis of Gabor filter bank. *Pacific-Rim Conference on Multimedia* (pp.719-730)

Le, R., Nguyen, M., Yan, W. (2021) Training a convolutional neural network for transportation sign detection using synthetic dataset. *International Conference on Image and Vision Computing New Zealand.*

Le, R., Nguyen, M., Yan, W., Nguyen, H. (2021) Augmented reality and machine learning incorporation using YOLOv3 and ARKit. *Applied Sciences.*

Le, R., Nguyen, M., Yan, W. (2021) A novel curtain style pictorial marker for enhancing augmented reality experiences. *International Conference on Image and Vision Computing New Zealand*.

Le, R. (2022) *Synthetic Data Annotation for Enhancing the Experiences of Augmented Reality Application Based on Machine Learning (PhD Thesis).* Auckland University of Technology, New Zealand.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., & Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. IEEE Conference on Computer Vision and Pattern Recognition, pp. 4681-4690.

Lee, H., Lee, J., Kim, H., Cho, B., & Cho, S. (2018). Deep-neural-network-based sinogram synthesis for sparse-view CT image reconstruction. IEEE Transactions on Radiation and Plasma Medical Sciences, pp. 109-119.

Li, C., Yan, W. (2021) Braille recognition using deep learning. *International Conference on Control and Computer Vision.*

Li. C. (2022) *Special Character Recognition Using Deep Learning. Master's Thesis* Auckland University of Technology, New Zealand.

Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., & Jurafsky, D. (2017). Adversarial learning for neural dialogue generation. arXiv:1701.06547v5

Li, P. (2018) *Rotation Correction for License Plate Recognition*. Master's Thesis, Auckland University of Technology, New Zealand.

Li, P., Nguyen, M., Yan, W. (2018) Rotation correction for license plate recognition. *International Conference on Control, Automation and Robotics*.

Li, R., Nguyen, M., Yan, W. (2017) Morse codes enter using finger gesture recognition. *International Conference on Digital Image Computing: Techniques and Applications*.

Li, Y., Ming, Y., Zhang, Z., Yan, W., Wang, K. (2021) An adaptive ant colony algorithm for autonomous vehicles global path planning. *International Conference on Computer Supported Cooperative Work in Design*.

Liang, B., Jia, X. X., & Lu, Y. (2021). Application of adaptive image restoration algorithm based on sparsity of block structure in environmental art design. Complexity.

Ling, H., Wang, L., Zou, F., Yan, W. (2011) Fine-search for image copy detection based on local affine-invariant descriptor and spatial dependent matching. Multimedia Tools and Applications 52 (2), 551-568.

Ling, H., Cheng, H., Ma, Q., Zou, F., Yan, W. (2011) Efficient image copy detection using multi-scale fingerprints. *IEEE Multimedia*, 19, 60–69

Ling, H., Feng, H., Zou, F., Yan, W., Lu, Z. (2010) A novel collusion attack strategy for digital fingerprinting. *International Workshop on Digital Watermarking*, 224-238.

Liou, C. Y., Cheng, C. W., Liou, J. W., & Liou, D. R. (2012). Autoencoder for polysemous word. International Conference on Intelligent Science and Intelligent Data Engineering, pp. 458-465.

Liu, G., Reda, F. A., Shih, K. J., Wang, T. C., Tao, A., & Catanzaro, B. (2018). Image inpainting for irregular holes using partial convolutions. European Conference on Computer Vision, pp. 85-100.

Liu, J., & Jung, C. (2021). Facial image inpainting using attention-based multi-level generative network. Neurocomputing, pp. 95-106.

Liu, K., Kang, G., Zhang, N., & Hou, B. (2018). Breast cancer classification based on fully-connected layer first convolutional neural networks. IEEE Access, pp. 23722-23732.

Liu, M., Yan, W. (2022) Masked face recognition in real-time using MobileNetV2. *ACM ICCCV.*

Liu, S., Bai, W., Zeng, N., & Wang, S. (2019). A fast fractal based compression for MRI images. IEEE Access, pp. 62412-62420.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. (2017). A survey of deep neural network architectures and their applications. Neurocomputing, pp. 11-26.

Liu, X. (2019) *Vehicle-related Scene Understanding Using Deep Learning.* Master's Thesis, Auckland University of Technology, New Zealand.

Liu, X., Yan, W. (2020) Vehicle-related scene segmentation using CapsNets. *International Conference on Image and Vision Computing New Zealand.*

Liu, X., Yan, W. (2021) Traffic-light sign recognition using Capsule network. *Springer Multimedia Tools and Applications.*

Liu, Z., Yan, W., Yang, B. (2018) Image denoising based on a CNN model. International Conference on Control, Automation and Robotics.

Liu, J., Ling, H., Zou, F., Yan, W., Lu, Z. (2012) Digital image forensics using multi-resolution histograms. *Crime Prevention Technologies and Applications for Advancing Criminal.*

Liu, Z. (2018) *Comparative Evaluations of Image Encryption Algorithms.* Master's Thesis, Auckland University of Technology, New Zealand.

Lu, J., Nguyen, M., Yan, W. (2018) Pedestrian detection using deep learning. IEEE AVSS.

Lu, J. (2021) *Deep Learning Methods for Human Behavior Recognition*. PhD Thesis. Auckland University of Technology, New Zealand.

Luo, Z., Nguyen, M., Yan, W. (2022) Kayak and sailboat detection based on the improved YOLO with Transformer. *ACM ICCCV.*

Luo, Z., Nguyen, M., Yan, W. Sailboat detection based on automated search attention mechanism and deep learning models. *International Conference on Image and Vision Computing New Zealand.*

Ma, X. (2020) *Banknote Serial Number Recognition Using Deep Learning.* Master's Thesis, Auckland University of Technology, New Zealand.

Ma, X., Yan, W. (2021) Banknote serial number recognition using deep learning. *Springer Multimedia Tools and Applications*.

Mehtab, S., Yan, W. (2021) FlexiNet: Fast and accurate vehicle detection for autonomous vehicles-2D vehicle detection using deep neural network. *International Conference on Control and Computer Vision*.

Mehtab, S., Yan, W. (2022) Flexible neural network for fast and accurate road scene perception. *Multimedia Tools and Applications.*

Mehtab, S. Yan, W., Narayanan, A. (2022) 3D vehicle detection using cheap LiDAR and camera sensors. *International Conference on Image and Vision Computing New Zealand.*

Mehtab, S. (2022) *Deep Neural Networks for Road Scene Perception in Autonomous Vehicles Using LiDARs and Vision Sensors.* PhD Thesis, Auckland University of Technology, New Zealand.

Meng, Q., Catchpoole, D., Skillicom, D., & Kennedy, P. J. (2017). Relational autoencoder for feature extraction. International Joint Conference on Neural Networks, pp. 364-371.

Neff, R., & Zakhor, A. (2002). Matching pursuit video coding. I. Dictionary approximation. IEEE Transactions on Circuits and Systems for Video Technology, pp. 13-26.

Nguyen, M., Yan, W. Temporal colour-coded facial-expression recognition using convolutional neural network. *International Summit Smart City 360°: Science and Technologies for Smart Cities.*

Niitsuma, M., Tomita, Y., Yan, W., Bell, D. (2018) Towards musicologist-driven mining of handwritten scores. *IEEE Intelligent Systems*.

Niitsuma, M., Tomita, Y., Yan, W., Bell, D. (2011) Classifying Bach's handwritten C-Clefs. *International Society for Music Information Retrieval Conference (ISMIR 2011)*.

Palubinskas, G. (2017). Image similarity/distance measures: What is really behind MSE and SSIM. International Journal of Image and Data Fusion, pp. 32-53.

Pan, C., Yan, W. (2018) A learning-based positive feedback in salient object detection. International Conference on Image and Vision Computing New Zealand.

Pan, C., Yan, W. (2020) Object detection based on saturation of visual perception. *Multimedia Tools and Applications*, 79 (27-28), 19925-19944.

Pan, C., Liu, J., Yan, W., Zhou, Y. (2021) Salient object detection based on visual perceptual saturation and two-stream hybrid networks. *IEEE Transactions on Image Processing.*

Pfau, D., & Vinyals, O. (2016). Connecting generative adversarial networks and actor-critic methods. arXiv:1610.01945v3

Popoola, O. P., & Wang, K. (2012). Video-based abnormal human behavior recognition—A review. IEEE Transactions on Systems, Man & Cybernetics: Part C - Applications & Reviews, pp. 865-878.

Qi, J., Nguyen, M., Yan, W. (2022) Waste classification from digital images using ConvNeXt. Pacific-Rim Symposium on Image and Video Technology.

Qin, X., Chen, W., Shen, Q., Jiang, J., & Feng, G. (2017). Image inpainting: a contextual consistent and deep generative adversarial training approach. IAPR Asian Conference on Pattern Recognition, pp. 588-593.

Qin, Z., Yan, W. (2021) Traffic-sign recognition using deep learning. *International Symposium on Geometry and Vision.*

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434v2

Razzak, M. I., Naz, S., & Zaib, A. (2018). Deep learning for medical image processing: Overview, challenges and the future. Classification in BioApps, pp. 323-350.

Rebecq, H., Ranftl, R., Koltun, V., & Scaramuzza, D. (2019). Events-to-video: Bringing modern computer vision to event cameras. IEEE/CVF Conference on Computer

Vision and Pattern Recognition, pp. 3857-3866.

Ren, Y., Nguyen, M., Yan, W. (2018) Real-time recognition of series seven New Zealand banknotes. *International Journal of Digital Crime and Forensics* (IJDCF) 10 (3), 50-66.

Richard, M. M., & Chang, M. Y. (2001). Fast digital image inpainting. International Conference on Visualization, Imaging and Image Processing, pp. 106-107.

Rudin, L. I., Osher, S., & Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena, pp. 259-268.

Sainath, T. N., Kingsbury, B., Saon, G., Soltau, H., Mohamed, A. R., Dahl, G., & Ramabhadran, B. (2015). Deep convolutional neural networks for large-scale speech tasks. Neural Networks, pp. 39-48.

Sainath, T. N., Mohamed, A. R., Kingsbury, B., & Ramabhadran, B. (2013). Deep convolutional neural networks for LVCSR. IEEE International Conference on Acoustics. Speech and Signal Processing, pp. 8614-8618.

Santana, E., & Hotz, G. (2016). Learning a driving simulator. arXiv preprint arXiv:1608.01230.

Sara, U., Akter, M., & Uddin, M. S. (2019). Image quality assessment through FSIM, SSIM, MSE and PSNR: A comparative study. Journal of Computer and Communications, pp. 8-18.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. Annals of Statistics, pp. 1875-1897.

Shen, D., Xin, C., Nguyen, M., Yan, W. (2018) Flame detection using deep learning. *International Conference on Control, Automation and Robotics.*

Shen, H., Kankanhalli, M., Srinivasan, S., Yan, W. (2004) Mosaic-based view enlargement for moving objects in motion pictures. *IEEE ICME'04*.

Shen, J., Yan, W., Miller, P., Zhou, H. (2010) Human localization in a cluttered space using multiple cameras. *IEEE International Conference on Advanced Video and Signal Based Surveillance.*

Shen, Y., Yan, W. (2019) Blind spot monitoring using deep learning. *International Conference on Image and Vision Computing New Zealand*.

Song, C., He, L., Yan, W., Nand, P. (2019) An improved selective facial extraction model for age estimation. *International Conference on Image and Vision Computing New Zealand.*

Song, Z., Tomasetto, F., Yan, W., Li, Y., et al. (2022) Enabling breeding selection for biomass in slash pine using UAV-based imaging. *Plant Phenomics.*

Shinde, P. P., & Shah, S. (2018). A review of machine learning and deep learning applications. International Conference on Computing Communication Control and Automation, pp. 1-6.

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., & Webb, R. (2017). Learning from simulated and unsupervised images through adversarial training. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2107-2116.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, pp. 484-489.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, pp. 1929-1958.

Suganuma, M., Ozay, M., & Okatani, T. (2018). Exploiting the potential of standard convolutional autoencoders for image restoration by evolutionary search. International Conference on Machine Learning, pp. 4771-4780.

Sulam, J., & Elad, M. (2016). Large inpainting of face images with trainlets. IEEE Signal Processing Letters, pp. 1839-1843.

Tong, D., Yan, W. (2022) Visual watermark identification from the transparent window of currency by using deep learning. *Applications of Encryption and Watermarking for Information Security.*

Vaishaya, R., Javaid, M., Khan, H. I., & Haleem, A. (2020). Artificial intelligence (AI) applications for COVID-19 pandemic. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, pp. 337-339.

Vargas, J. A., Esquivel, J. Z., & Tickoo, O. (2021). Introducing region pooling learning. International Conference on Pattern Recognition, pp. 714-724.

Wang, C., Jiang, Y., Wang, K., & Wei, F. (2021). A field-programmable gate array system for sonar image recognition based on convolutional neural network. Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering, pp. 1808-1818.

Wang, G., Wu, X., Yan, W. (2017) The state-of-the-art technology of currency identification: A comparative study. International Journal of Digital Crime and Forensics 9 (3), 58-72.

Wang, H., Yan, W. (2022) Face detection and recognition from distance based on deep learning. *Aiding Forensic Investigation Through Deep Learning and Machine*

Learning Framework. IGI Global.

Wang, J., Yan, W., Kankanhalli, M., Jain, R., Reinders, M. (2003) Adaptive monitoring for video surveillance. International Conference on Information, Communications and Signal Processing.

Wang, J., Kankanhalli, M., Yan, W., Jain, R. (2003) Experiential sampling for video surveillance. *ACM SIGMM International Workshop on Video surveillance* (pp.77-86).

Wang, J., Yan, W. (2016) BP-neural network for plate number recognition. International *Journal of Digital Crime and Forensics* (IJDCF) 8 (3), 34-45.

Wang, J. (2016) *Event-driven Traffic Ticketing System*. Master's Thesis, Auckland University of Technology, New Zealand.

Wang, J., Bacic, B., Yan, W. (2018) An effective method for plate number recognition. *Multimedia Tools and Applications,* 77 (2), 1679-1692.

Wang, L., Yan, W. (2021) Tree leaves detection based on deep learning. *International Symposium on Geometry and Vision.*

Wang, P., Patel, V. M., & Hacihaliloglu, I. (2018). Simultaneous segmentation and classification of bone surfaces from ultrasound using a multi-feature guided CNN. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 134-142.

Wang, W., Huang, Y., Wang, Y., & Wang, L. (2014). Generalized autoencoder: A neural network framework for dimensionality reduction. IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 490-497.

Wang, X., Yan, W. (2019) Human gait recognition based on self-adaptive hidden Markov model. *IEEE/ACM Transactions on Biology and Bioinformatics*.

Wang, X., Yan, W. (2019) Cross-view gait recognition through ensemble learning. *Neural Computing and Applications.*

Wang, X., Yan, W. (2019) Gait recognition using multichannel convolutional neural networks. *Neural Computing and Applications.*

Wang, X., Yan, W. (2019) Multi-perspective gait recognition based on ensemble learning. *Springer Neural Computing and Applications*.

Wang, X., Yan, W. (2019) Human gait recognition based on frame-by-frame gait energy images and convolutional long short-term memory. *International Journal of Neural Systems.*

Wang, X., Yan, W. (2020) Non-local gait feature extraction and human identification.

Springer Multimedia Tools and Applications.

Wang, X., Yan, W. (2020) Cross-view gait recognition through ensemble learning. *Neural computing and applications* 32 (11), 7275-7287.

Wang, X., Yan, W. (2022) Human identification based on gait manifold. *Applied Intelligence.*

Wang, Y. (2021) *Colorizing Grayscale CT Images of Human Lung Using Deep Learning.* Master's Thesis, Auckland University of Technology, New Zealand.

Wang, Y., Yan, W. (2022) Colorising grayscale CT images of human lungs using deep learning methods. *Springer Multimedia Tools and Applications.*

Weir, J., Lau, R., Yan, W. (2012) Digital image splicing using edges. *International Journal of Digital Crime and Forensics* (pp.176-187)

Wei, T., Li, Q., Liu, J., Zhang, P., & Chen, Z. (2020). 3D face image inpainting with generative adversarial nets. Mathematical Problems in Engineering.

Wiriyathammabhum, P., Summers-Stay, D., Fermüller, C., & Aloimonos, Y. (2016). Computer vision and natural language processing: Recent approaches in multimedia and robotics. ACM Computing Surveys, pp. 1-44.

Wu, J. M., Li, Z., Herencsar, N., Vo, B., & Lin, J. C. (2021). A graph-based CNN-LSTM stock price prediction algorithm with leading indicators. Multimedia Systems, pp. 1-20.

Wu, Y., Singh, V., & Kapoor, A. (2020). From image to video face inpainting: Spatial-temporal nested GAN (STN-GAN) for usability recovery. IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2396-2405.

Xiang, Y., Yan, W. (2021) Fast-moving coin recognition using deep learning. *Springer Multimedia Tools and Applications.*

Xiao, B., Nguyen, M., Yan, W. (2021) Apple ripeness identification using deep learning. *International Symposium on Geometry and Vision.*

Xin, C. (2020) *Detection and Recognition for Multiple Flames Using Deep Learning.* Master's Auckland University of Technology, New Zealand.

Xin, C., Nguyen, M., Yan, W. (2020) Multiple flames recognition using deep learning. *Handbook of Research on Multimedia Cyber Security*, 296-307.

Xing, J., Yan, W. (2021) Traffic sign recognition using guided image filtering. *International Symposium on Geometry and Vision.*

Xing, J., Nguyen, M., Yan, W. (2022) The improved framework of traffic sign recognition by using guided image filtering. *Springer Nature Computer Science.*

Xing, J. (2022) *Traffic Sign Recognition from Digital Images Using Deep Learning.* Master's Thesis, Auckland University of Technology, New Zealand.

Xing, J., Nguyen, M., Yan, W. (2022) Traffic sign recognition from digital images by using deep learning. Pacific-Rim Symposium on Image and Video Technology.

Xie, J., Xu, L., & Chen, E. (2012). Image denoising and inpainting with deep neural networks. International Conference on Neural Information Processing Systems, pp. 341-349.

Xu, A., Chen, J., Zhang, P., & Wu, J. (2011). Information fusion method for ocular aberrations measurement based on subjective visual compensation. Optik, pp. 1240-1244.

Yan, W., Kankanhalli, M. (2002) Erasing video logos based on image inpainting. *IEEE International Conference on Multimedia and Expo*, 521-524.

Yan, W., Kankanhalli, M. (2002) Detection and removal of lighting & shaking artifacts in home videos. *ACM International Conference on Multimedia*, 107-116.

Yan, W., Kankanhalli, M., Wang, J., Reinders, M. (2003) Experiential sampling for monitoring. *ACM SIGMM Workshop on Experiential Telepresence*, 70-72.

Yan, W., Kankanhalli, M. (2003) Colorizing infrared home videos. *International Conference on Multimedia and Expo.*

Yan, W., Kankanhalli, M., Wang, J. (2005) Analogies-based video editing. *Multimedia Systems* 11 (1), 3-18.

Yan, W., Wang, J., Kankanhalli, M. (2005) Automatic video logo detection and removal. *Multimedia Systems* 10 (5), 379-391.

Yan, W., Kankanhalli, M. (2007) Multimedia simplification for optimized MMS synthesis. *ACM Transactions on Multimedia Computing, Communications, and Applications.*

Yan, W., Kankanhalli, M. (2009) Cross-modal approach for Karaoke artefacts correction. *Handbook of Multimedia for Digital Entertainment and Arts*, 197-218.

Yan, W., Kieran, D., Rafatirad, S., Jain, R. (2011) A comprehensive study of visual event computing. Multimedia Tools and Applications 55 (3), 443-481.

Yan, W., Chambers, J. (2013) An empirical approach for digital currency forensics. *IEEE International Symposium on Circuits and Systems* (ISCAS), 2988-2991.

Yan, W., Chambers, J., Garhwal, A. (2014) An empirical approach for currency identification. *Multimedia Tools and Applications* 74 (7).

Yan, W., Kankanhalli, M. (2015) Face search in encrypted domain. Pacific-Rim Symposium on Image and Video Technology, 775-790.

Yan, W., Liu, F. (2015) Event analogy-based privacy preservation in visual surveillance. Pacific-Rim Symposium on Image and Video Technology, 357-368.

Yan, W. (2019) *Introduction to Intelligent Surveillance: Surveillance Data Capture, Transmission, and Analytics*. Springer London.

Yan, W. (2021) *Computational Methods for Deep Learning: Theoretic, Practice and Applications*. Springer London.

Yan, Z., Li, X., Li, M., Zuo, W., & Shan, S. (2018). Shift-net: Image inpainting via deep feature rearrangement. European Conference on Computer Vision, pp. 1-17.

Yang, Y., Cheng, Z., Yu, H., Zhang, Y., Cheng, X., Zhang, Z., & Xie, G. (2021). MSE-Net: generative image inpainting with multi-scale encoder. Visual Computer, pp. 1-13.

Yang, Z., Hu, Z., Salakhutdinov, R., & Berg-Kirkpatrick, T. (2017). Improved variational autoencoders for text modeling using dilated convolutions. International Conference on Machine Learning, pp. 3881-3890.

Yin, X., Goudriaan, J. A., Lantinga, E. A., Vos, J. A., & Spiertz, H. J. (2003). A flexible sigmoid function of determinate growth. Annals of botany, pp. 361-371.

Yoo, H. J. (2015). Deep convolution neural networks in computer vision: A review. IEIE Transactions on Smart Processing and Computing, pp. 35-43.

Younas, F., Usman, A., Yan, W. (2022) A deep ensemble learning method for colorectal polyp classification with optimized network parameters. *Applied Intelligence.*

Yu, G., Sapiro, G., & Mallat, S. (2010). Image modeling and enhancement via structured sparse model selection. IEEE International Conference on Image Processing, pp. 1641-1644.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2018). Generative image inpainting with contextual attention. IEEE Conference on Computer Vision and Pattern Recognition, pp. 5505-5514.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2019). Free-form image inpainting with gated convolution. IEEE/CVF International Conference on Computer Vision, pp. 4471-4480.

Yu, L., Zhang, W., Wang, J., & Yu, Y. (2017). SeqGAN: Sequence generative adversarial nets with policy gradient. AAAI Conference on Artificial Intelligence.

Yu, Z. (2021) *Deep Learning Methods for Human Action Recognition*. Master's Thesis, Auckland University of Technology, New Zealand.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. European Conference on Computer Vision, pp. 818-833.

Zeiler, M. D., Krishnan, D., Taylor, G. W., & Fergus, R. (2010). Deconvolutional networks. IEEE Computer Vision and Pattern Recognition, pp. 2528-2535.

Zeng, Y., Fu, J., Chao, H., & Guo, B. (2019). Learning pyramid-context encoder network for high-quality image inpainting. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1486-1494.

Zhang, H., & Li, T. (2020). Semantic face image inpainting based on generative adversarial network. Youth Academic Annual Conference of Chinese Association of Automation, pp. 530-535.

Zhang, L. (2020) *Virus Identification from Digital Images Using Deep Learning*. Master's Thesis, Auckland University of Technology, New Zealand.

Zhang, L., Yan, W. (2020) Deep learning methods for virus identification from digital images. *International Conference on Image and Vision Computing New Zealand*.

Zhang, Q. (2018) *Currency Recognition Using Deep Learning*. Master's Thesis, Auckland University of Technology, New Zealand.

Zhang, Q., Yan, W. (2018) Currency detection and recognition based on deep learning. *IEEE International Conference on Advanced Video and Signal Based Surveillance*.

Zhang, Q., Yan, W., Kankanhalli, M. (2019) Overview of currency recognition using deep learning. *Journal of Banking and Financial Technology,* 3 (1), 59–69.

Zhang, R., Quan, W., Wu, B., Li, Z., & Yan, D. M. (2020). Pixel-wise dense detector for image inpainting. Computer Graphics Forum, pp. 471-482.

Zhang, X., Wang, X., Shi, C., Yan, Z., Li, X., Kong, B., & Mumtaz, I. (2022). De-GAN: Domain embedded gan for high quality face image inpainting. Pattern Recognition.

Zhang, Y., Gan, Z., & Carin, L. (2016). Generating text via adversarial training. NIPS Workshop on Adversarial Training, pp. 21-32.

Zhang, Y. (2016) *A Virtual Keyboard Implementation Based on Finger Recognition.* Master's Thesis, Auckland University of Technology, New Zealand.

Zhang, Y., Yan, W., Narayanan, A. (2017) A virtual keyboard implementation using finger recognition. *International Conference on Image and Vision Computing New Zealand.*

Zhao, K. (2021) *Fruit Detection Using CenterNet*. Master's Thesis, Auckland University of Technology, New Zealand.

Zhao, K., Yan, W. (2021) Fruit detection from digital images using CenterNet. *International Symposium on Geometry and Vision.*

Zheng, C., Cham, T. J., & Cai, J. (2019). Pluralistic image completion. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1438-1447.

Zheng, K., Yan, Q., Nand, P. (2017) Video dynamics detection using deep neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Zhou, L., Yan, W., Shu, Y., Yu, J. (2018) CVSS: A cloud-based visual surveillance system. *International Journal of Digital Crime and Forensics* (IJDCF) 10 (1), 79-91.

Zhu, Y., Yan, W. (2022) Ski fall detection from digital images using deep learning. *ACM ICCCV.*

Zhu, Y., Yan, W. (2022) Image-based storytelling using deep learning. *ACM ICCCV*.

Zhu, Y., Yan, W. (2022) Parasite detection from digital images using deep learning methods. Machine Learning and AI Techniques in Interactive Medical Image Analysis, IGI Global.

Zhu, Y., Yan, W. (2022) Traffic sign recognition based on deep learning. *Multimedia Tools and Applications.*

Zou, J., Yan, W., Ding, W., Qi, D. (2001) A novel image texture substitution with shading effect. *Journal of Computer Research and Development* 38 (11), 1327-1330.