A DENOISING GENERATIVE ADVERSARIAL NETWORK BASED ON ENHANCED FEATURE MAPPING OF DATA AUGMENTATION FOR IMAGE SYNTHESIS

CHEN LI

UNIVERSITI SAINS MALAYSIA

2024

A DENOISING GENERATIVE ADVERSARIAL NETWORK BASED ON ENHANCED FEATURE MAPPING OF DATA AUGMENTATION FOR IMAGE SYNTHESIS

by

CHEN LI

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

May 2024

ACKNOWLEDGEMENT

After three and a half years, my doctoral studies at USM are coming to an end. I pursued my master's degree at USM in 2017 and will be graduating with a PhD in 2024. The support, encouragement, and companionship from those around me have been the biggest motivation for me to complete my degree. As i reflect on this period of study, I would like to express my sincere gratitude to all those who have helped and supported me. First of all, i would like to express my gratitude to my supervisor Dr. Chan Huah Yong for his invaluable assistance in my research work during my PhD study. Dr. Chan provided me with the freedom to choose my research direction and offered many valuable suggestions. Additionally, Dr. Chan maintained open communication with me on a regular basis. I would like to express my gratitude towards the positive influence that his attitude towards academics and life has had on me, allowing me to calmly face the pressure of studies. Additionally, I would like to thank the members of the panel, Dr. Nur Intan Raihana Ruhaiyem and Dr. Tan Tien Ping for their constructive comments and guidance throughout the proposal and research review stages of my thesis. At the same time, I would like to express my gratitude to my former classmates: Yaru Wang, Le Zhou, Yuhong Yang, and KT Wong. Yaru's suggestion largely influenced my decision to pursue my PhD degree. Le Zhou and Yuhong Yang provided me with invaluable assistance in both my studies and personal life, while KT Wong helped me resolve many of my research-related doubts. Lastly, I would like to thank my parents for their unwavering support and care, which has been my strongest source of motivation.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF SYMBOLS

z	Latent code from Prior Distribution
E	Mathematical Expectation
p_x	Probability Distribution of x
$\overline{z}$	Latent Code with Noise
λ_1	The Hyperparameter of Noise Penalty Term
λ_2	The Hyperparameter of Standard Deviation Term
$\overline{x}$	Mean Value of x
α	Learning Rate
β_1	The Hyperparameter of Adam
β_2	The Hyperparameter of Adam
θ_g	The Parameter of Generator Network
θ_e	The Parameter of Encoder Network
θ_d,	The Parameter of Discriminator Network
U	Standard Normal Distribution
∇	The Computation of Gradient

LIST OF ABBREVIATIONS

GANs	Generative Adversarial Networks
DNFM	Denoising Feature Mapping
CelebA	Celebrity Faces Attribute
LSUN	Large-scale Image Dataset using Deep Learning with Humans in the Loop
IS	Inception Score
FID	Fréchet Inception Distance Score
AWT	Amazon Mechanical Turk
VAE	Variational Autoencoder
ELBO	Evidence Lower Bound
JS	Jensen-Shannon
ICA	Independent Component Analysis
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
RBM	Restricted Boltzmann Machine
RNN	Recurrent Neural Networks
PCA	Principle Component Analysis
MSE	Mean Square Error
KL	Kullback-Leibler
CVAE	Conditional Variational Autoencoder
NICE	Non-liner Independent Components Estimation
GLOW	Generative Flow with Invertible
MMGAN	Minimax Generative Adversarial Networks
NSGAN	Non-saturating Generative Adversarial Networks

DAE	Denoising Autoencoder
CTAE	Contractive Autoencoder
DFM	Denoising Feature Mapping
BEGAN	Boundary Equilibrium Generative Adversarial Networks
CoGAN	Coupled Generative Adversarial Networks
MRGAN	Mode Regularized Generative Adversarial Networks
SAGAN	Self-Attention Generative Adversarial Networks
BGAN	Boundary-Seeking Generative Adversarial Networks
COCO	Conditional Coordinating
AGE	Adversarial Generator-Encoder Networks
VAT	Virtual Adversarial Training
PPGN	Plug and Play Generative Networks
SWGAN	Sliced Wasserstein Generative Adversarial Networks
SPADE	Spatially-Adaptive Normalization
VQVAE	Vector Quantized Variational Autoencoder
EMD	Earth Mover Distance
WGAN-GP	Wasserstein Generative Adversarial Networks with Gradient Penalty
DRAGAN	Deep Regret Analytic Generative Adversarial Networks
VGAN	Variational Generative Adversarial Networks
SNGAN	Spectrum Normalized Generative Adversarial Networks
SGD	Stochastic Gradient Descent
MMDGAN	Maximum Mean Discrepancy Generative Adversarial Networks
RGAN	Relativistic Generative Adversarial Networks
RaGAN	Relative Average Generative Adversarial Networks
LSGAN	Least Square Generative Adversarial Networks

MNIST	National Institute of Standards and Technology
LapGAN	Laplacian Pyramid of Adversarial Networks
ProGAN	Progressive Growth Generative Adversarial Networks
TTUR	Two Time-Scale Update Rule
CGAN	Conditional Generative Adversarial Networks
ACGAN	Auxiliary Classifier Generative Adversarial Networks
EBGAN	ENERGY-BASED Generative Adversarial Networks
NP	Noise Penalty
BN	Batch Normalization
Relu	Linear Rectification Function

LIST OF APPENDICES

RANGKAIAN BERTENTANGAN GENERATIF PENGURANGAN KEBISINGAN BERDASARKAN PEMETAAN CIRI AUGMENTASI DATA YANG DIPERTINGKATKAN UNTUK SINTESIS IMEJ

ABSTRAK

Rangkaian bertentangan generatif (GAN) telah menjadi topik penyelidikan yang penting dalam pembelajaran mendalam untuk sintesis imej. GAN boleh menghasilkan hasil yang pelbagai dan berkualiti tinggi melalui kerjasama antara penjana dan diskriminator. Walau bagaimanapun, membina model GAN yang teguh dan stabil kekal sebagai cabaran penting. Penyelidikan terdahulu telah cuba meningkatkan GAN asal dengan menggunakan pelbagai algoritma untuk mengukur perbezaan antara pengagihan data, melaksanakan struktur rangkaian yang berbeza, atau menggabungkannya dengan struktur lain untuk mencapai hasil yang lebih baik. Walau bagaimanapun, penambahbaikan ini selalunya terhad kepada satu perspektif. Kertas kerja ini memperkenalkan GAN Pemetaan Ciri Denoising (DNFM-GAN), varian GAN yang meningkatkan kestabilan latihan model dengan menambah baik komponen penjana dan diskriminator. Khususnya, keupayaan penjana dipertingkatkan dengan menambahkan data dengan hingar sebagai input tambahan. Ini memerlukan penjana untuk mempelajari cara menjana imej daripada data yang rosak separa, yang membawa kepada perwakilan yang lebih baik yang dipelajari daripada data tersebut. Untuk memastikan kestabilan dan keteguhan penjana, adalah penting untuk meminimumkan turun naik yang disebabkan oleh kehilangan penjana. Selain itu, menjana dua jenis data, G(z) dan G(z+noise) boleh meningkatkan kesukaran diskriminasi untuk diskriminasi apabila digabungkan dengan data sebenar. Selain itu, GAN tradisional sering menghadapi isu bahawa Jensen-Shannon Divergence

mengukur perubahan dalam jarak pengedaran secara tidak tepat, menjadikannya sukar untuk mengoptimumkan penjana. Untuk membimbing penjana dengan lebih baik agar sesuai dengan pengedaran data sebenar, kajian ini mencadangkan DNFM-GAN, yang menggunakan helah pemetaan ciri yang dipertingkatkan untuk mengemas kini penjana. Ini membolehkan penjana mendapatkan maklumat kecerunan yang lebih lancar, menghasilkan kestabilan latihan yang lebih baik. Eksperimen yang dijalankan pada set data menara CelebA dan LSUN menunjukkan bahawa DNFM-GAN boleh menghasilkan hasil yang memuaskan tanpa mengalami keruntuhan mod. Berbanding dengan ImprovedDCGAN, VAEGAN, dan WGAN-GP, model ini mencapai skor FID yang lebih rendah sebanyak 15.963 pada CelebA dan 12.956 pada menara LSUN, dan skor Inception yang lebih tinggi sebanyak 9.742 pada ClebA dan 8.225 pada menara LSUN pada resolusi 128×128. Dalam penilaian kualitatif berdasarkan AWT, DNFM-GAN mencapai peratusan tertinggi iaitu 51% dan 57%. Ini menunjukkan bahawa imej yang dijana oleh DNFM-GAN kelihatan unggul secara visual daripada perspektif manusia.

A DENOISING GENERATIVE ADVERSARIAL NETWORK BASED ON ENHANCED FEATURE MAPPING OF DATA AUGMENTATION FOR IMAGE SYNTHESIS

ABSTRACT

Generative adversarial networks (GANs) have become a significant research topic in deep learning for image synthesis. GANs can produce diverse and high-quality results through the collaboration between the generator and discriminator. However, building a robust and stable GANs model remains a significant challenge. Previous research has attempted to enhance the original GANs by utilizing various algorithms to measure divergence between data distributions, implementing different network structures, or combining them with other structures to achieve better results. But these improvements were often limited to a single perspective. This research introduces the Denoising Feature Mapping GAN (DNFM-GAN), a GAN variant that enhances the stability of the model's training by improving both the generator and discriminator components. Specifically, the generator's ability is enhanced by adding data with noise as an extra input. This requires the generator to learn how to generate images from partially damaged data, leading to better representations learned from the data. To ensure the generator's stability and robustness, it is important to minimize the volatility caused by generator loss. Additionally, generating two types of data, $G(z)$ and $G(z + noise)$ can increase the difficulty of discrimination for the discriminator when combined with real data. Moreover, traditional GANs often encounter the issue that Jensen-Shannon Divergence inaccurately measures changes in distribution distances, making it difficult to optimize the generator. To better guide the generator to fit the real data distribution, DNFM-GAN employs an enhanced feature mapping trick to

update the generator. This allows the generator to obtain smoother gradient information, resulting in improved training stability. Experiments conducted on the CelebA and LSUN towers datasets demonstrated that DNFM-GAN can produce satisfactory results without experiencing mode collapse. Compared to ImprovedDCGAN, VAEGAN, and WGAN-GP, this model achieved a lower FID score of 15.963 on CelebA and 12.956 on LSUN towers, and a higher Inception score of 9.742 on CelebA and 8.225 on LSUN towers at 128×128 resolution. In the qualitative assessment based on AWT, DNFM-GAN achieved the highest percentage with 51% and 57%. This suggests that images generated by DNFM-GAN appears superior results from a human perspective.

CHAPTER 1

INTRODUCTION

1.1 Research Background

With the recent advancements in deep learning, it has made significant achievements in the tasks of computer vision, especially image classification, image synthesis, etc. Generally, the task of image classification based on machine learning needs to predict data labels from input images. With recent advances in machine learning and artificial intelligence techniques (especially deep learning models), it achieves a great success that sometimes reach or even surpass human performance, such as in visual object recognition and object detection or image segmentation. As mentioned before, the task of image classification is to predict image label based on input image features. Specifically, the essence of this training process is to learn the conditional probability $p(y|x)$, that is, the output of these models should be the probability of image label y under the condition of given image sample x. Ultimately, the decision boundary is used to distinguish between different types of images. Nowadays, as long as there is enough training data for these kinds of tasks, the prediction accuracy rate can be relatively high. Many notable models such as Support Vector Machine (Boser et al., 1995), Alexnet (Krizhevsky et al., 2017), Resnet (He et al., 2015), Mobilenet (A. Howard et al., 2019; A. G. Howard et al., 2017; Sandler et al., 2019) can be classified into this category.

However, as American theoretical physicist Richard Frynman pointed out, "what I cannot create, I do not understand". Creating data itself may be a better indicator of understanding the data than data classification. Image synthesis based on deep learning is more difficult than discriminative tasks, which currently requires the help of generative models to produce richer information, such as complete images with certain

details and changes. Generative models focus on the sample distribution and model the data distribution. To be specific, it includes both latent variable models and supervised models. Where the latent variable model requires the model to be able to generate higher-dimensional results that conform to the real data distribution from the lower-dimensional code. These high dimensional results can be structured data, such as images, videos or audios, etc. For example, the hidden variable model can be defined as $p(x,z)$, where x is the real image itself, and z is the hidden variable, the process of image synthesis is to transfer code z from a 100-dimensional hidden variable to a 64×64 resolution image. Eventually, the model will learn the distribution of the real data that can produce images which are similar to the real data. Basically, this method requires the model to have a strong ability to match the objective distribution. The other is supervised learning models, which needs to consider the information of the image label as an extra constraint during the training process. So that, the supervised learning model needs to learn probability density $p(x,y)$. For example, y is a label of Arabic numeral 6, and x is a generated image under the y label, When the model generates images, it not only needs to consider the quality of the generated images, but also needs to consider the constraints of label information on the results. At present, there is extensive research on both models and both have a wide range of applications.

Nowadays, the advancement of image synthesis technology can not only facilitate the further development of traditional computer vision and graphics tasks, but also the related applications which are closely to people's lives are becoming increasingly mature. For example, people can combine image synthesis with text to control the generated images that conform to the current description or control the local features (Figure 1.2) of the generated facial image (Choi et al., 2018; Karras et al., 2019), such as change the facial expression, beard, hairstyle which is shown in Figure 1.2.

Moreover, many applications, such as the task of discriminative models, often require a large number of training sets, but obtaining these training set data is expensive and time-consuming, and image synthesis models can provide a way to solve this problem very well. Furthermore, the image synthesis model can do style transfer (Zhu et al., 2020), such as converting ordinary paintings into paintings of well-known painters (Figure 1.1), which also took a lot of time in the past by traditional way. To sum up, image synthesis is an important research direction in the field of computer vision, and it has significant research value from perspectives of both research and applications.

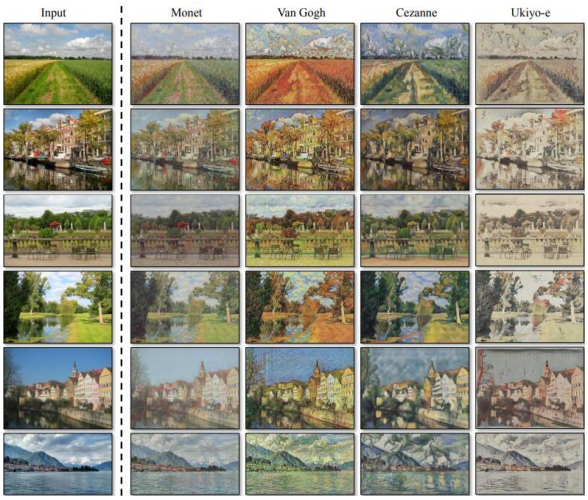

Figure 1.1 The Example of Style Transfer

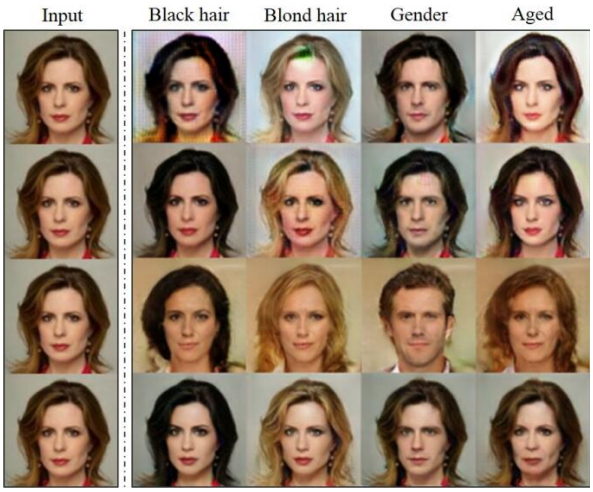

Figure 1.2 The Example of Feature Control

1.2 The Current Research Works

The research on image synthesis has been going on for a long time. The early study is generally based on the relationship between image pixels. These kinds of models often need to manually design features, which is relatively simple to implement and does not need to consume a lot of calculations. However, it is only suitable for images with simple content and lower resolution, once the content of image is complex, the effect will not work very well.

The development of machine learning has brought new opportunities for image synthesis technology. During this period, many models based on machine learning have made progress in image synthesis tasks, but the architecture of these models are generally shallow with a fixed model structure, and deal with specialized problems, so there are certain limitations in the capabilities and application scenarios of these models.

In recent years, image synthesis focusses on process of generating new images from a set of input parameters, the goal of image synthesis in this stage is to create realistic images that are not present in the original dataset, such as creating new faces from a dataset of faces, or creating new landscapes from a dataset of images of landscapes. the combination of deep neural network and probabilistic graphical model has produced many new models. For instance, the Autoregressive model (Oord, Kalchbrenner, & Kavukcuoglu, 2016; Oord, Kalchbrenner, Vinyals, et al., 2016) is a kind of linear prediction, which uses the linear combination of random variables at several moments in the past to describe the linear regression model of random variables at a certain moment in the future. The network structure of this model can adopt recurrent neural network or convolutional neural network based on different task requirements. Besides, another model is called Variational autoencoder (VAE)

(Kingma & Welling, 2013) which the structure is similar to that of autoencoder. The encoder and decoder structures remain unchanged, but the encoder of VAE does not directly output latent code, but outputs two codes m and σ, which represents the mean and variance of distribution. Intuitively, VAE can be regarded as a normal autoencoder adding noise to the input. After reconstruction process, the original image can be obtained. Another model is called Flow-based models (Dinh et al., 2015, 2017) that use normalizing flows to transform a simple latent variable distribution (such as a standard Gaussian) into a target distribution. The training of this kind of model is faster than auto-regressive model, and the optimization goal is the maximum likelihood estimation instead of optimizing the Evidence Lower Bound (ELBO) like the VAE, so the training is easy to implement. Basically, Flow-based models have been used for a variety of applications, including image synthesis, super-resolution, and generative design. They have also been used for tasks such as denoising, inpainting, and style transfer. However, Flow-based models have some limitations, it is difficult to model complex distributions. In addition, Calculating the Jacobian Matrix can be costly and hardly handle discrete variables or non-differentiable operations. Besides, the generated samples sometimes lack diversity and have inferior quality. Currently, the best quality of generated images is achieved by Generative Adversarial Network (GANs) (I. J. Goodfellow et al., 2014) which are currently mainly used in image synthesis, text generation, style transfer and other directions. Especially in the field of image synthesis, the achievements of GAN are impressive. GANs is different from previous models in that it does not directly estimate the probability density of the target distribution, but relies on its specific model structure to judge the difference between the distribution of generated data and real data. Basically, GAN consists of two important components which are the generator and the discriminator. Both structures are neural networks. The main task of the generator is to

generate high dimensional data through the lower dimensional input code z from prior distribution, the objective of generator is to "fool" the discriminator as much as possible, that means the generated image $G(z)$ can be considered by discriminator to be a real data instead of a fake data. While, the role of the discriminator is to judge whether the data is real data or the data generated by the "generator" which works like a binary classifier. The training process can be defined as follows: where x indicates the input data of discriminator, and the output $D(x)$ represents the probability that x is real data. If it is 1, it means discriminator has 100% confidence that x is the real data. Meanwhile, if the output is 0, it means that it cannot be regarded as real data. In this way, generator and discriminator constitute a dynamic confrontation (Minmax gaming). As the training (confrontation) progresses, the data generated by generator is getting closer to the real data, and the level of discriminator's identification data is also getting enhanced. In an ideal state, generator can produce data that is enough to "confuse the real". But for discriminator, it is difficult to determine whether the data generated by the generator is real or not, so $D\big(G(z)\big)$ equals to 0.5. After training, it is possible to get a generative model that can be used to generate “real data”. Although the quality of the images generated by GANs is very high, due to the differences in the tasks of the generator and the discriminator, the training of GANs is difficult to maintain stability, which is the current direction for researchers to improve GANs.

Through the review of image synthesis technology, it can be found that the development of this field has made great progress, and it is also one of the hotspots of computer vision research. However, ensuring the high quality, diversity, and robustness of the generated images remains a significant challenge for the models to overcome.

1.3 Research Problems

Image synthesis refers to the generation of new images through computer algorithms. It can be used in various applications such as image restoration, image generation, image style transfer, etc. GAN plays an important role in image synthesis as it can learn the data distribution and generate high quality images. The generator generates synthetic images by learning the statistical characteristics of the data, while the discriminator evaluates the similarity between the generated images and real images. The training process of GAN can be summarized as a confrontation game between the generator and the discriminator. The generator tries to generate realistic images to fool the discriminator, while the discriminator tries to distinguish between generated images and real images. Through this adversarial training, the generator gradually improves the quality of the generated images to the point where the discriminator cannot accurately distinguish between generated images and real images. Therefore, GAN is one of the commonly used methods in image synthesis. It can generate high-quality, realistic images and is widely used in fields such as computer vision, image processing, and artistic creation.

Despite the great success of GANs, it still has some problems that cannot be ignored. Firstly, the generator needs to map the data from the low-dimensional distribution to the high-dimensional distribution and there is no constraint on generator to constraint the generation space that cause insufficient robustness of the generator (Radford et al., 2016), which badly affects the generated results. So, how to make the generator more robust is one of the research problems of this paper. In order to better explain the problem, the experiment of original GAN training was trained 150k steps on the Fashion_Minist dataset and illustrated the change of generator loss. It could be found that although the overall loss shows a downward trend, there is big volatility of

loss of generator which suddenly become larger. Meanwhile, images with this larger loss are extracted at steps 96600 and 112600, it was found that the quality of these images is poor with many strange textures. This reflects the lack of robustness of generator, and it is tough for generator to learn a good intermediate representation of data. The reason for this phenomenon is largely due to the training mechanism of GANs. GAN does not need to directly estimate the probability density of the target distribution. It lacks a clear training goal such as optimizing ELBO like VAE. Hence, the generator is too arbitrary in training and lacks robustness. If there is no constraint on the generator, the quality of the generated images will be badly affected. This phenomenon is shown by figures 1.3 and 1.4.

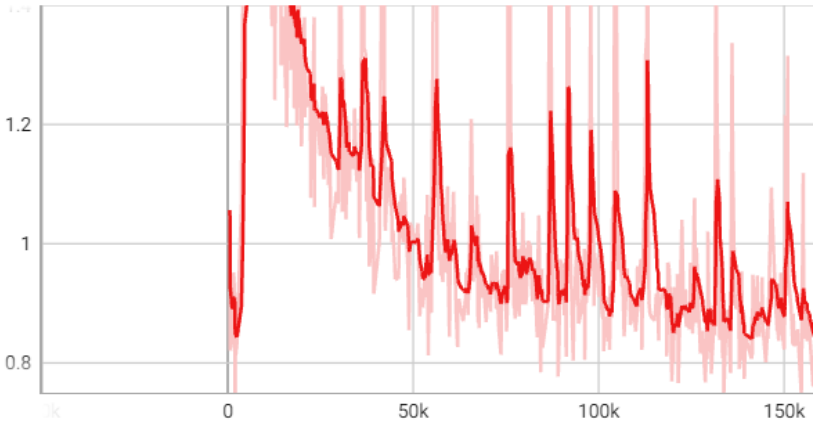

Figure 1.3 The Loss of Generator Training on Fashion_Minist

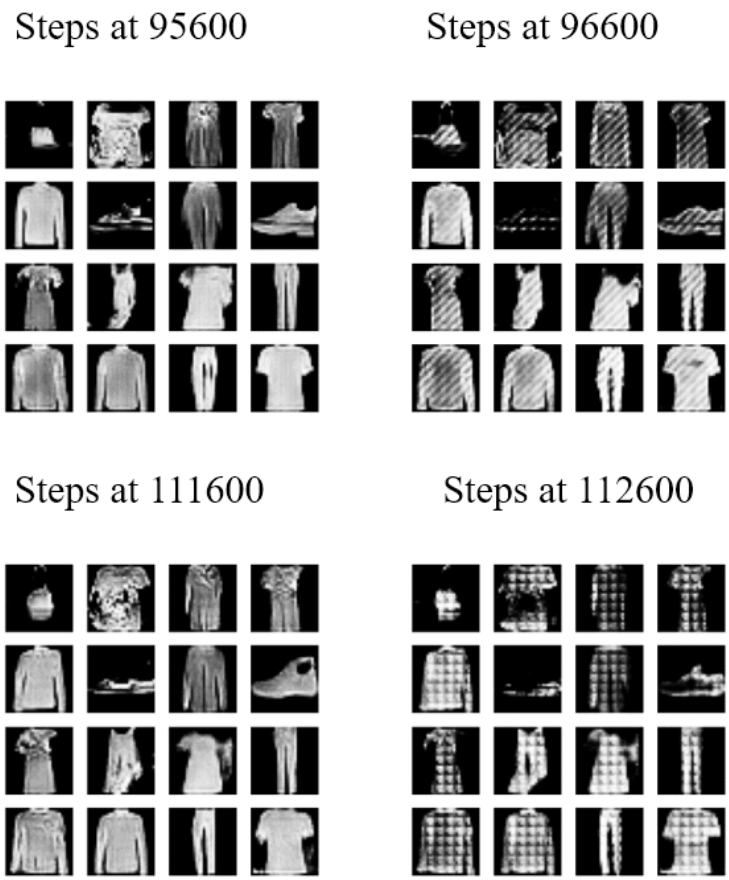

Figure 1.4 The Generated Results at Step 96600 and 112600

In addition, GAN requires generator and discriminator alternate training. But it is difficult for the generator and discriminator to converge at the same time. Most of the deep model training takes gradient descent optimization algorithm to reach at optimized point of the loss function and eventually reach at the local minimum or saddle point. Generative adversarial neural network requires both generator and discriminator to keep an equilibrium situation during the training game (I. Goodfellow et al., 2014). For the above reasons, the optimization methods with a same gradient direction may reduce the gradient for generator, but increase gradient of discriminator. Even sometimes the two sides of the game finally reach at equilibrium point, they are constantly offsetting each other's progress. This leads to that generator and discriminator cannot reach at the optimal convergence point simultaneously. In addition, the difference in the difficulty

of the two components also makes it difficult for them to converge synchronously. The discriminator only needs to complete data discrimination, which is relatively easy to train, while generator needs to complete the mapping from low-dimensional data to high-dimensional data, and the task is relatively more complicated. Because the latent code z is usually sampled from normal distribution or gaussian distribution randomly at the beginning of training, fake data distribution is too far away from the real data distribution and the image is just a low-dimensional manifold in a high-dimensional space, so it is difficult to have overlapping areas between different distributions or even if there is any overlap, it can be ignored. This cause the problem of gradient vanishing of generator, that is when the ability of discriminator is very powerful, the loss of the discriminator quickly equals to 0, which cannot provide a reliable path to continue to update the gradient of the generator, causing the gradient of the generator to vanish (Salimans et al., 2016). Therefore, increasing the discrimination difficulty of the discriminator so that it is not easy for discriminator to distinguish between real and fake samples in the early stages of training can avoid the loss of discriminator being too small, thus improving the stability of the model.

Moreover, the optimization goal of generator is to minimize the distance between the target distribution and the generated distribution, and this goal is often difficult to achieve in training. To be specific, the loss function of discriminator from original GAN is defined as:

$$V(G,D) = E_{x\sim pdata}\left[LogD(x)\right] + E_{x\sim pG}\left[Log\big(1 - D(x)\big)\right] \tag{1.1}$$

Where, E indicates mathematical expectation, and $x\sim pdata$, $x\sim pG$ represents data from real distribution and generated distribution respectively. And discriminator needs to maximize the equation below during training process:

$$D^* = arg\max_D V(D,G) \tag{1.2}$$

Obviously, discriminator should maximize the value of $D(x)$ if $x \sim pdata$, and minimize the value of $D(x)$ if $x \sim pG$. Integral transformation of formula (1.1) can get:

$$V(G,D) = E_{x\sim pdata}\left[LogD(x)\right] + E_{x\sim pG}\left[Log\big(1-D(x)\big)\right]$$

$$= \int_x P_{data}(x) logD(x)dx + \int_x P_G(x)\log\big(1-D(x)\big)\,dx$$

$$= \int_x \left[P_{data}(x) logD(x) + P_G(x)\log\big(1-D(x)\big)\right]dx \tag{1.3}$$

Assuming $D(x)$ is any function, then only need to find $D(x)$=$D^*(x)$, when $D^*(x)$ reaches the maximum value, the formula 1.3 obtains the maximum value. Let $P_{data}(x)$ = a, $P_G(x)$ = b, $D(x) = D$, the formula 1.3 can be:

$$f(D) = alog(D) + blog(1-D) \tag{1.4}$$

The extreme point of formula 1.4 can be get by:

$$\frac{df(D)}{dD} = a \times \frac{1}{D} + b \times \frac{1}{1-D} \times (-1)$$

$$D^* = \frac{a}{a+b}$$

$$D^* = \frac{P_{data}(x)}{P_{data}(x) + P_G(x)} \tag{1.5}$$

So that, when $D(x) = D^*$,

$$\max_D V(G, D^*) = E_{x \sim P_{data}} \left[\frac{P_{data}(x)}{P_{data}(x) + P_G(x)} \right] + E_{x \sim P_G} \left[\frac{P_{data}(x)}{P_{data}(x) + P_G(x)} \right]$$

$$= \int_x P_{data}(x) log \frac{P_{data}(x)}{P_{data}(x) + P_G(x)} + \int_x P_G(x) log \frac{P_{data}(x)}{P_{data}(x) + P_G(x)}$$

$$= -2log2 + KL(P_{data} || \frac{P_{data} + P_G}{2} + KL(P_G || \frac{P_{data} + P_G}{2})$$

$$= -2log2 + 2JSD(P_{data} \parallel P_G) \quad (1.6)$$

As can be seen from above formula (1.6) that maximizing D^* is maximizing the Jensen-Shannon divergence (JSD) between P_G and P_{data}. Similarly, for the generator, it is necessary to minimize the distance between P_G and P_{data} which can be described by $G^* = arg \min_G \max_D V(G, D)$.

However, there are some problems with adopting Jensen–Shannon divergence (JS) to measure the difference of two distributions that this metric does not accurately measure the distance between two distributions, since according to the nature of JS divergence, even if the distance between two distributions is closer than the previous iteration, if two distributions do not overlap, the distance between them will always be $log2$ (Arjovsky et al., 2017) (Figure 1.5). Since the images are only the low-dimensional manifolds in a high-dimensional space, and during training, all data is sampled from P_G and P_{data}, there is almost no overlap between them, or even if there is overlap, the overlapping area can be ignored. Under this condition, the generator will

not be able to obtain useful gradient information to update, resulting in the failure of the entire training. Therefore, how to accurately measure the difference between the generated distribution and the real distribution to guide the generator to produce better quality data is also a problem.

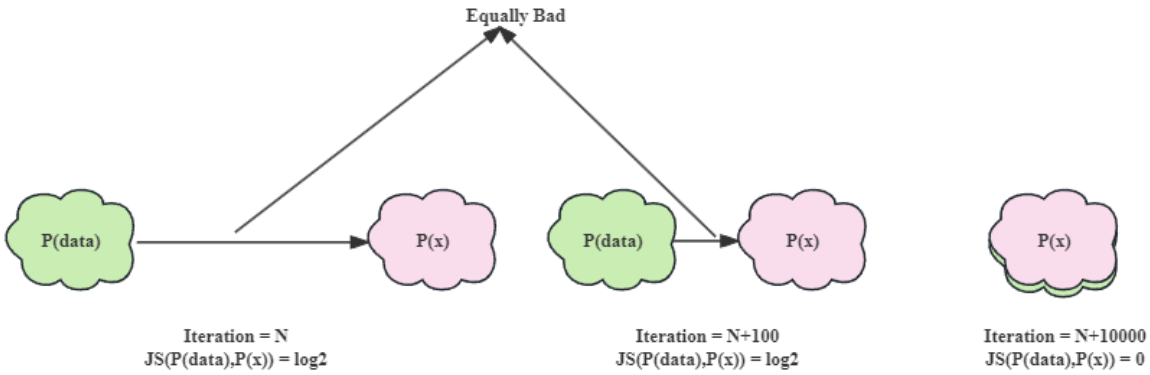

Figure 1.5 The Shortcoming of JS Divergence

According to the analysis above, the research problems are defined as follows:

1. There are no constraints on the generator, which can cause volatility in loss values and make it difficult for the generator to learn a good intermediate representation of the data.

2. The discriminator is able to easily distinguish between generated data and real data during training, causing the discriminator's loss to approach 0.

3. The Jensen-Shannon Divergence is not an accurate measure of the divergence between the real and generated distributions, which can cause unstable training for the discriminator and hinder the generator's convergence.

1.4 Research Questions

According to the issues identified in this study, there are three main challenges that need to be addressed. Firstly, the generator lacks constraints, leading to inadequate robustness, difficulty in acquiring a good representation of data, and the inability to stably produce high-quality images. Secondly, the high-dimensional space presents a challenge for the image manifold. This causes a lack of overlap between real and generated data, making it easier for the discriminator to distinguish between the two and resulting in a small loss. Finally, the use of JS divergence as a measure for the distance between different distributions is not accurate, thus guiding the generator towards unsatisfactory results.

Therefore, the focus of this research is to answer three significant questions. The first question is how to introduce appropriate constraints to the generator to increase its robustness and enable it to generate high-quality images consistently. The second question is how to address the issue of the lack of overlap between real and generated data in the high-dimensional space to improve the performance of the discriminator. The third question pertains to the development of accurate measures for the distance between different distributions to guide the generator towards generating high-quality images. By answering these questions, the research aims to overcome the challenges identified and improve the quality of the generated images.

1. What constraint can enhance the robustness of generator? How to design this constraint?

2. How to increase the difficulty of discrimination for discriminator? How to combine this concept into loss function?

3. How to design the optimization objective of generator to accurately measure the distance change between real data and fake data?

1.5 Research Objective

This research sheds light on the drawbacks of GANs. Presently, GANs encounter several problems that need to be addressed. Firstly, one significant issue entails the absence of any constraint on the generator in traditional GANs. This results in the volatility of loss values and lower quality of produced samples. Therefore, it is vital to impose constraints on the generator to limit the generation space and enhance its robustness. Implementing bounds on the generated samples can enable the generator to produce more controllable and predictable results by preventing it from generating unrealistic samples. This can be achieved through various techniques, including regularization or gradient penalty methods. These methods enable the generator to learn a more accurate and stable representation of the domain, which makes it less sensitive to input perturbations, leading to better results. Additionally, maintaining a balance between the generator and discriminator is crucial for optimal performance of GANs. In conclusion, by adding constraints to the generator, GANs can effectively resolve the current issues and improve the quality and robustness of the generated data, resulting in more accurate and reliable output for a variety of applications.

Next, due to the task difference between generator and discriminator. Discriminator can easily distinguish between real data and generated data in the stage of training, and the value of loss will be very small. This will make it difficult for generator to obtain effective gradient information, and it is necessary to increase the

difficulty of discrimination of the discriminator. Under this condition, it is not too easy for discriminator to distinguish between real data and fake data.

Furthermore, the optimization method of generator needs to be improved, according to the characteristics of JS divergence, using JS divergence to measure the distance of different distributions has certain limitations, if the generator just simply maximizes the value of output form discriminator, this will cause the problem of gradient vanishing, and generator cannot obtain smooth gradient information to update. In general, the objectives of this work are as follows:

1. To enhance the stability of generator, the structure with an additional decoder will be included to reconstruct data with noise, this will enable the generator to learn a more stable and effective intermediate representation of the data.

2. To design a mechanism that can increase the difficulty of discrimination of the discriminator, so that the loss of the discriminator will not be too small which can be better to provide gradient information to generator.

3. To propose the enhanced feature mapping optimization method for generator to get more accurate convergence of generator and stable training process, this will improve the results of image synthesis.

1.6 Research Scope

The primary focus of this study is to design a new variant of Generative Adversarial Networks (GANs) to get the high-quality results of image synthesis, with the objective of achieving consistent production of high-quality and varied images.

Specifically, the research is focused on addressing three key aspects of GANs: namely, reducing the instability of the generator's loss, enhancing the discriminator's ability to accurately discriminate between real and generated data, introducing a more effective approach to measuring the divergence 0 between distributions of real data and fake data. By improving these critical aspects of GANs, this research aims to push the boundaries of what is currently possible in the field of generative image synthesis, providing more sophisticated and refined outputs with greater consistency and diversity.

It also designs corresponding experiments to verify the effectiveness of these three improvements. Concretely, the dataset CelebA (Z. Liu et al., 2015) and LUSN towers (Yu et al., 2016) are used to train the proposed model. These two datasets contain rich face and tower samples respectively, which are common for training generative models (Denton et al., 2015; Larsen et al., 2016; Roth et al., 2017; Zhang, Xu, et al., 2019). The resolution of generated images by this proposed model is 128×128, which is used for evaluation and compare to other models, They are WGAN-GP (Gulrajani et al., 2017), ImprovedDCGAN (Salimans et al., 2016), and VAEGAN (Larsen et al., 2016). Furthermore, Since there is currently no unified standard for evaluating generative models, in order to evaluate the model more comprehensively, this study takes two objective metrics Inception score (IS) (Barratt & Sharma, 2018) and Fréchet Inception Distance (FID) (Heusel et al., 2018) as the criteria for evaluation of the performance of model, these two metrics are commonly used in current research on generative models. In addition, this research also takes an evaluation standard Amazon Mechanical Turk (AMT) (Shaham et al., 2019) based on human subjectivity as a supplementary evaluation criterion. However, other datasets and evaluation metrics are not considered in this study.

1.7 Research Contributions

Generating realistic images is one of the key goals of image synthesis tasks. The advantage of GAN in image synthesis is its ability to learn the statistical characteristics of the data and generate images similar to the training data. It has the ability to generate high-quality, realistic images, and can achieve more precise control and diversity by adjusting the model's structure and training strategy. This research proposes a variant of GANs model called DNFM-GAN. Compared with the traditional GAN, the following improvements are made. First, it learns more robust data features through a denoising structure. To be specific the model adds an additional encoder, which can convert $G(z + nosise)$ into $\bar{z}$ and generator needs to minimize the difference between z and $\bar{z}$. This can help generator to learn more robust representation from data. It can be found from the experimental results that this method can improve the ability of generator. Moreover, these generated data with noise are put into discriminator together with real data and generated data for discrimination. This method increases the difficulty of discriminator's discrimination, so that the loss of discriminator will not be too small. Finally, DNFM-GAN adopts discrepancy of mean and standard deviation as a measure of distribution in feature mapping. The mean value can reflect the feature center, while the standard deviation can describe the feature spread. If the mean and standard deviation of the two data is very close, they are more likely to come from the same distribution. This trick can help generator to generate data closer to real data than using mean as metric alone. Experiments have shown that this method is effective, and no mode collapse was found in the experiments. The contributions of this paper as shown in the following:

1. A denoising structure that can improve the robustness of generator for image synthesis on learning good intermediate representation of data which alleviate loss volatility,

2. A mechanism of data augmentation that can increase the difficulty of discrimination to avoid the gradient vanishing of the discriminator,

3. An enhanced feature mapping which is based on the discrepancy of mean and standard deviation of distributions that can better measure the difference of distributions which help generator to produce more realistic images.

1.8 Thesis Outline

The first chapter provides an overview of the fundamental principles of image composition and the importance of investigating this field. Moreover, it delivers a brief examination of existing methods that are commonly employed in this area of study. Building on these foundational concepts, the research queries are introduced, which defines the aims of the study and summarizes the principal contributions it aims to achieve.

In the second chapter, the evolution and limitations of image synthesis techniques are described in detail, including early techniques and techniques based on deep learning models, especially generative adversarial networks. According to the research objectives, this chapter reviews the application of techniques which can improve the robustness of model. At the same time, it also refers to the improved method which can enhance the stability of model training, including the improvements about model structures, objective functions and the optimization methods. The models

introduced in these documents are sorted out, and their advantages and limitations are illustrated.

Inspired by literature review, chapter three proposes a GAN model name DNFM-GAN, which describes in detail how to solve the proposed research problems. the model architecture is illustrated and explained in detail, the improved loss function, and training steps of the entire model are also specifically presented. The hyperparameters of this model is shown in this chapter also.

In chapter 4, the proposed method is subjectively and objectively evaluated on different dataset which shows that this model is stable in training and can generate high quality images, it also be compared with other previous models.

Finally, the conclusion, limitations and future work of this research are discussed in chapter 5.

CHAPTER 2

LITERATURE REVIEW

This chapter collects and organizes relevant literatures about image synthesis. First of all, it reviews the development of image synthesis technology, mainly including image synthesis models based on pixels and textures, image synthesis models based on shallow structure, and image synthesis models based on deep learning. These parts mainly introduce the background of image synthesis models and their respective characteristics. Then, the improvements of GANs models, which is closely related to this research problems are discussed in terms of robust structure and improvements of model training stability, the strengths and limitations of these models are also discussed. Finally, it introduces the current evaluation criteria for image synthesis.

2.1 The Development of Image Synthesis Models

Image synthesis refers to the process of generating new images using computer algorithms. The development of image synthesis technology can be traced back to the 1980s. As mentioned before, image synthesis technology often needs to be linked with generative models. In the early days, limited by the calculation power of devices and the development stage of computer vision technology, image synthesis was mainly based on pixel relationships and texture structures to generate some simple line segments and regular components. Strictly speaking, the model at this stage cannot be regarded as a real image synthesis model, since it just simply stitches or transfers some parts of the image. With the development of machine learning, image synthesis models began to combine with machine learning technology, resulting in some models such as Independent Component Analysis model (ICA) (Aapo et al., 2001), Gaussian mixture model (GMM) (Permuter et al., 2003), Hidden Markov Model (HMM) (Starner &

Pentland, 1995), Restricted Boltzmann Machine (RBM) (Salakhutdinov & Hinton, 2009) and so on. These models need to use manual feature learning to establish relevant features for the original data. Compared with the previous models, these models can generate the texture and simple objects better, but the depth of these models is often shallow and the structure of the network is fixed for specific tasks. Although they do not cost much computation power, it is not capable to process images with complex structures, rich content, and high-resolution.

In recent years, the development of deep learning and representation learning provides effective means for image synthesis technology (Hinton & Salakhutdinov, 2006). During this period, the combination of deep learning models and probabilistic graphical models gave birth to classic models such as Autoregressive model, Variational Auto Encoder (VAE), Flow-based Model and GANs. These models often have more complex model structures with deep layers and can better learn the distribution of the target data, compared with the previous model, they made great strides in image synthesis tasks.

2.1.1 The Early Models

The research on image synthesis has been going on for a long time, and the early research on image synthesis is generally based on the relationship between image pixels or texture information. But these methods only realize some simple tasks, such as background color change (Smith & Blinn, 1996), image patching (Criminisi et al., 2003), image extrapolating (Efros & Leung, 1999), simple image texture synthesis (Heeger & Bergen, 2014), and image enhancement (Chakkarwar & Shandilya, 2013) which are shown in Figure 2.1. Strictly speaking, these methods cannot be regarded as real image synthesis. Meanwhile, these kinds of models are based on manual ways, researchers are required to have domain knowledge, which is difficult to implement. Although it does

not need to consume a lot of computation, it is only suitable for generating images with simple content and low resolution. If the content of the image is complex, the effect will not work very well.

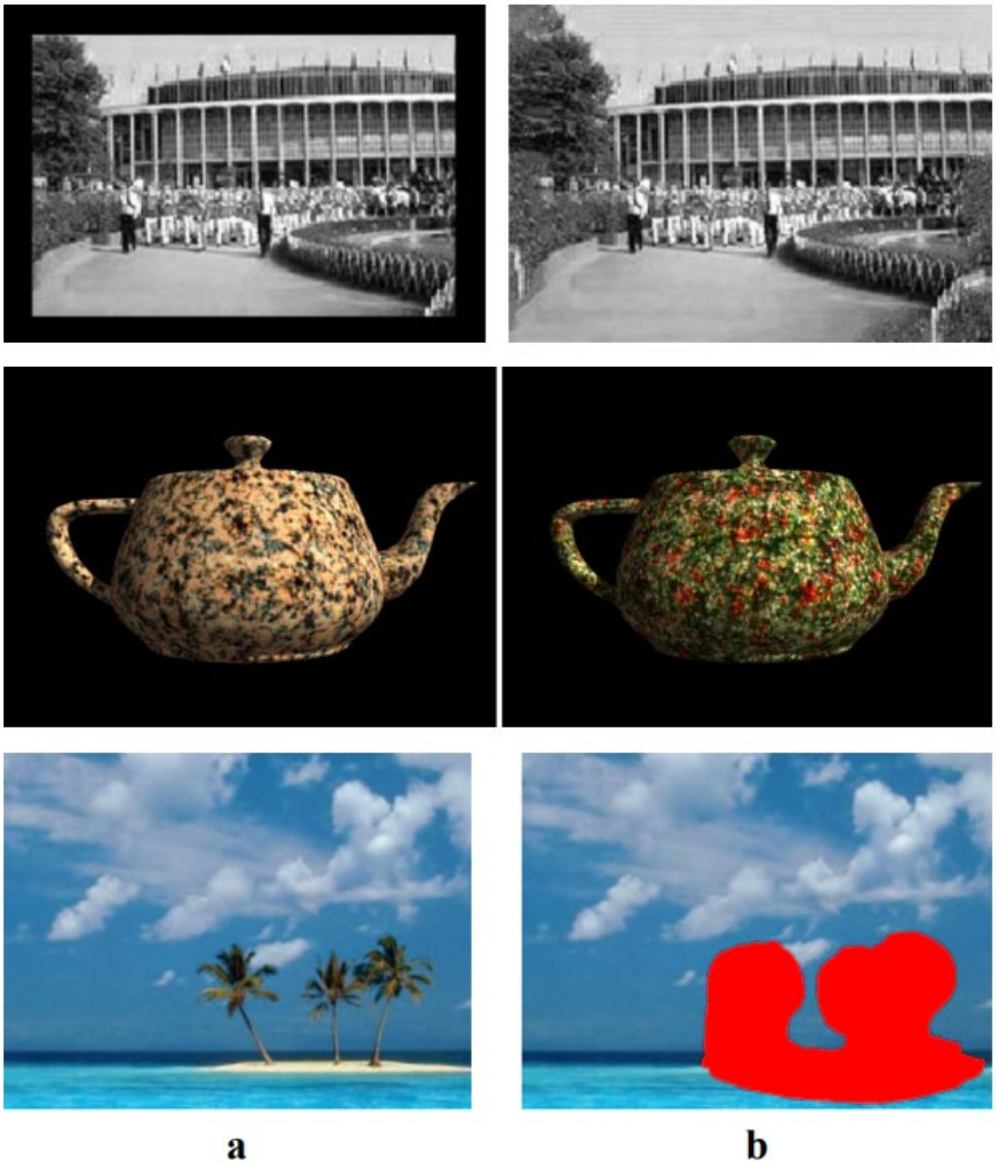

Figure 2.1 The Examples of Image Synthesis by Early Models

The advances of machine learning and feature representation learning provide a new way for image synthesis. Some novel models such as Independent Component Analysis model (Hyvärinen et al., 2004), Gaussian mixture model (Permuter et al., 2003; Theis et al., 2012; Xu & Jordan, 1996), and Hidden Markov Model (Starner & Pentland,

1995) have been developed. Since these models have a relatively fixed structure, they cannot fit more complex image distributions, so it is basically suitable for images with regular, simple objects. Subsequent models such as Markov Random Field (Harrison, 2001; Ranzato et al., 2010) and Restricted Boltzmann Machine (Hinton, 2002; Hinton & Salakhutdinov, 2006; Salakhutdinov & Hinton, 2009) have achieved great success in image synthesis compared with previous model, but they often need to consume a lot of computation to repeatedly calculate the Markov chain, and the expression of nonlinearity of these models is relatively not enough. At that time, there was still a lack of a means to handle image synthesis with high-resolution, complex content task.

2.1.2 Autoregressive Model

The development of deep learning has made it possible to train multi-layer neural network models. Since that time, the capability of neural network model has made great progress and the image synthesis technology has opened a new development mileage with the help of deep learning. Pixel Recurrent Network (PixelRNN) (Oord, Kalchbrenner, & Kavukcuoglu, 2016) was enlightened by the idea of Recurrent Neural Networks (RNN) (Lipton et al., 2015) in sequence generation and proposed Pixel Recurrent Network (PixelRNN). More specifically, the aim of this study is to model a density distribution or likelihood function $p(x)$ of the data. It depends on chain rule to decompose log-likelihood into the product of one-dimensional distribution. Once the likelihood functions are defined, the training objective is to maximize the likelihood of the training data under this definition, pixel by pixel, until a complete image is generated. However, determining the optimal order of each pixel can be challenging. PixelCNN (Oord, Kalchbrenner, Vinyals, et al., 2016) has basically the same idea as PixelRNN, but took differences in network structure. Basically, PixelRNN or PixelCNN needs to use the chain rule to calculate the maximum likelihood function as below: