7-21-2022

An Information-theoretic analysis of generative adversarial networks for image restoration in physics-based vision

Xudong KANG

Terms of Use

AN INFORMATION-THEORETIC ANALYSIS OF GENERATIVE ADVERSARIAL NETWORKS FOR IMAGE RESTORATION IN PHYSICS-BASED VISION

KANG XUDONG

MPHIL

LINGNAN UNIVERSITY

2022

AN INFORMATION-THEORETIC ANALYSIS OF GENERATIVE ADVERSARIAL NETWORKS FOR IMAGE RESTORATION IN PHYSICS-BASED VISION

by
KANG Xudong
康旭東

A thesis
submitted in partial fulfillment
of the requirements for the Degree of
Master of Philosophy in Business

Lingnan University

2022

ABSTRACT

An Information-theoretic Analysis of
Generative Adversarial Networks
for Image Restoration in Physics-based Vision

by

KANG Xudong

Master of Philosophy

Image restoration in physics-based vision (such as image denoising, dehazing, and deraining) are fundamental tasks in computer vision that attach great significance to the processing of visual data as well as subsequent applications in different fields. Existing methods mainly focus on exploring the physical properties and mechanisms of the imaging process, and tend to use a deconstructive idea in describing how the visual degradations (like noise, haze, and rain) are integrated with the background scenes. This idea, however, relies heavily on manually engineered features and handcrafted composition models, which can be theories only in ideal conditions or hypothetical models that may involve human bias or fail in simulating true situations in actual practices. With the progress of representation learning, generative methods, especially generative adversarial networks (GANs), are considered a more promising solution for image restoration tasks. It directly learns the restorations as end-to-end generation processes using large amounts of data without understanding their physical mechanisms, and it also allows completing missing details / damaged information by involving external knowledge and generating plausible results with intelligent-level interpretation and semantics-level understanding of the input images. Nevertheless, existing studies that try to apply GAN models to image restoration tasks dose not achieve satisfactory performances compared with the traditional deconstructive methods. And there is scarcely any study or theory to explain how deep generative models work in relevant tasks.

In this study, we analyzed the learning dynamics of different deep generative models based on the information bottleneck principle and propose an information-theoretic framework to explain the generative methods for image restoration tasks. In which, we study the information flow in the image restoration models and point out three sources of information involved in generating the restoration results: (i) high-level information extracted by the encoder network, (ii) low-level information from the source inputs that retained, or pass directed through the skip connections, and, (iii) external information introduced by the learned parameters of the decoder network during the generation process.

Based on this theory, we pointed out that conventional GAN models may not be directly applicable to the tasks of image restoration, and we identify three key issues leading to their performance gaps in the image restoration tasks: (i) over-invested abstraction processes, (ii) inherent details loss, and (iii) imbalance optimization with vanishing gradient. We formulate these problems with corresponding theoretical analyses and provide empirical evidence to verify our hypotheses and prove the existence of these problems respectively.

To address these problems, we then proposed solutions and suggestions including optimizing network structure, enhancing details extraction and accumulation with network modules, as well as replacing measures of training objectives, to improve the performances of GAN models on the image restoration tasks. Ultimately, we verify our solutions on bench-marking datasets and achieve significant improvement on the baseline models.

DECLARATION

I declare that this is an original work based primarily on my own research, and I warrant that all citations of previous research, published or unpublished, have been duly acknowledged.

SIGNED

(KANG Xudong)

CERTIFICATE OF APPROVAL OF THESIS

AN INFORMATION-THEORETIC ANALYSIS OF
GENERATIVE ADVERSARIAL NETWORKS
FOR IMAGE RESTORATION IN PHYSICS-BASED VISION

by

KANG Xudong

Master of Philosophy

Panel of Examiners :

SIGNED

(Prof. ZHANG Yue) (Chairman)

SIGNED

(Dr TAO Xiaohui) (External Member)

SIGNED

(Prof. XIE Haoran) (Internal Member)

SIGNED

(Prof. DAI Hongning Henry) (Internal Member)

Chief Supervisor :

Prof. XIE Haoran

Co-supervisor :

Prof. WONG Man Leung

Approved for the Senate :

SIGNED

(Prof. MOK Ka Ho Joshua)
Chairman, Postgraduate Studies Committee

21 JUL 2022

Date

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Field, Scope & Topic of Research

1.1.1 Computer Vision

Visual data such as images and videos are often of higher dimensionality and contain richer information compared to other conventional types of data Gao et al. (2017). Whereas, humans can interpret even more from these data than the information contained in their pixel values, such as identifying objects involved in the images and predicting the motion of each individual, by utilizing our prior knowledge. This advanced perception in the human visual system based on learning processes has inspired the development of artificial intelligence (AI), where machines are also supposed to learn to analyze, extract abstract features, or recognize complicated patterns from the visual data more than just numerical processing Nixon and Aguado (2019).

Computer vision (CV) Nixon and Aguado (2019); Parker (2010); Szeliski (2010) is the field of artificial intelligence that studies how machines can work like human vision system to intelligently process this visual information and perform various complex tasks such as object detection Zhao, Zheng, Xu, and Wu (2019) and semantic segmentation Garcia-Garcia et al. (2018). With the rapid progress of this field, a growing number of promising methods and algorithms have been proposed to assist in the analysis and understanding of these visual data, which lights up more possibilities for various intelligent-level tasks and decision-making.

1.1.2 Physics-based Vision

Images and videos we observe are partial reflections of the 3D physical world after intricate processes of light. Physics-based vision Brand (1997); Wolff, Shafer, and Healey (1993) is a branch of computer vision that tries to estimate and recover these intricate properties of the captured scene by modeling the physical imaging process Y. Li, Fu, Liang, and Zheng (2019); Y. Yu (2021). Factors it studies include shading, reflectance, medium properties, weathering, et al. These scene properties attach great significance to the recovery of the real scene information during capturing and can be essential for the subsequent processing and analysis of these data.

A broad variety of tasks are involved in physics-based vision, such as shape recovery from shading R. Zhang, Tsai, Cryer, and Shah (1999), reflection separation Wan, Shi, Duan, Tan, and Kot (2017), and photometric stereo Ackermann and Goesele (2015). Numerous of them are directly associated with practical applications (such as scene depth estimation Bhoi (2019)),

while others can be essential pre-procedures for subsequent downstream tasks (like bad weather restoration for object detection Chengtao, Qiuyu, and Yanhua (2015); Gui et al. (2021); B. Li et al. (2018); S. Li et al. (2019); W. Yang, Tan, Wang, Fang, and Liu (2020a)). Moreover, many physics-based vision studies also inspire tasks in many other subjects and have significant influences on interdisciplinary research, including the simulation and mathematical modeling in optical physics, the digital synthesis in computer graphics, as well as the analysis and interpretation of medical images et al Metaxas (2012).

1.1.3 Image Restoration

Image restoration is a typical set of tasks in physics-based vision, which has long been an essential topic in computer vision research.

Images captured by optical sensors or devices (like cameras) are often visually degraded by various factors from both internal (such as noise, blur, aliasing, and compression artifact inside the camera) or external (such as rain, fog, haze, and other weather distortions) sources. These visual degradations are inevitable in the formation, transmission, and storage of image data Gunturk and Li (2018), due to the imperfection of devices and hardware Fan, Zhang, Fan, and Zhang (2019); Motwani, Gadiya, Motwani, and Harris (2004), or restricted by the actual observation scenarios (such as bad weather or natural atmospheric absorption Chengtao et al. (2015); Tan (2008)). In addition, a single captured image can only contain limited information about the observed scene: we can only capture and store the observed images in bitmap format (array of pixels) (technically, there is no way to take photos in vector graphics), which can only contain a limited number of pixel values, determined by the resolution of the sensors or capturing devices Farsiu, Robinson, Elad, and Milanfar (2004); Nasrollahi and Moeslund (2014).

Image restoration tasks are attempting to recover the missing / hidden information from a visually degraded input and obtain a clean / higher-quality image Gunturk and Li (2018). Relevant tasks include not only the removals of noise and various distortions Chengtao et al. (2015); Fan et al. (2019); S. Li et al. (2019); Motwani et al. (2004); W. Yang et al. (2020a) but also image inpainting Elharrouss, Almaadeed, Al-Maadeed, and Akbari (2020) (completing the missing contents of an image), image interpolation (such as super-resolution Farsiu et al. (2004); Nasrollahi and Moeslund (2014)) et al. These tasks are foundations of low-level vision, which can be helpful for the subsequent high-level tasks (such as identification and object detection) and benefit a wide range of applications and research in different fields.

In this thesis, we focus on the three specific tasks of image restoration in physics-based vision: image denoising, image dehazing, and image de-raining. They are the most common examples of image restoration and can often represent and reflect other tasks in image restoration.

Image Denoising

As the very basic component in digital communication, noise is unavoidably involved in all digital images during their acquisition, compression, and transmission processes, causing the reduction of visual quality and loss of image information. Image denoising, which is to remove this noise and recover the latent observation, has been studied for decades as one of the fundamental problems in image restoration Fan et al. (2019); Goyal, Dogra, Agrawal, Sohi, and Sharma (2020); Motwani et al. (2004); Tian et al. (2020).

Image Dehazing

The existence of particles in the air (like dust, aerosol, and water droplet) as well as the natural atmospheric absorption often reduce the visibility of images captured about the observed scene, which are commonly known as haze or fog according to the particles or degradation involved. Image dehazing and defogging try to remove these factors from the image in order to obtain a clean background scene Chengtao et al. (2015); Gui et al. (2021); B. Li et al. (2018).

Image Deraining

Similarly, images captured during bad weather conditions may be visually degraded by rain streaks, rain accumulations, and raindrops adherent to the camera, thus image deraining tasks attempt to remove these rain factors from the observed rainy image and try to recover a clean rain-free background as the outputsS. Li et al. (2019); W. Yang et al. (2020a).

1.2 Background & Recent Advances

1.2.1 Early Development of Image Restoration

The earliest methods for image restoration tasks Habibi (1972); Nahi (1972); Vastola and Poor (1984) do not consider them as physics-based vision problems, where some numeral operations in image processing or "statistical image enhancement" techniques are adopted in a bid to obtain good-looking images Gunturk and Li (2018).

It was later realized that: rather than generally processing, we need to extract/simulate these visual degradations so as to better remove them from the inputs and achieve more impressive restoration results. Therefore, early ideas in physics-based vision, like causal analysis Brand (1997), are applied to image restoration tasks, trying to explore the physical properties and understand the mechanisms behind the imaging processes. These methods intend to estimate the absolute values of specific physical quantities and to design appearance models that well-simulate the observations fully based on existing physics theories, mathematic derivations, geometry calculations, and relevant prior knowledge Shafer, Kanade, Klinker, and Novak (1990).

However, not all tasks in physics-based vision can be accurately simulated or modeled based on well-studied prior knowledge. In many cases, especially for image restoration tasks, their visual appearances or imaging processes often involve randomness, which cannot be precisely

modeled or predicted. Moreover, many tasks are subject to complex physical systems that can not be exactly described using existing prior knowledge or may be affected by various unknown factors that do not have generic laws to be founded upon. For instance, the Lambertian reflectance model Koppal (2020) has long been regarded as a solid theory to describe the diffuse reflections of a matte surface, but in practice, it fails when specularities exist Y. Yu (2021). Thus, the early idea of causal analysis above only tends to work on those tasks with relatively simple physical processes or with mature and complete theoretical bases (like illumination estimation Hordley (2006) and radiometric calibration Wyatt (2012)), but not on the image restoration.

As an improvement, more advanced approaches attempt to use a parameterized approximation idea. Based on the existing frameworks of physical models and theories, these methods introduce learnable parameters in their hypothetical models to allow variances in describing the phenomena for the randomness and unknown factors involved. And then, using the idea of statistical learning, these models try to approximate and simulate the complex systems by optimizing the models' parameters and fitting the results with certain amounts of data Y. Li et al. (2019); Y. Yu (2021). Famous examples include the Additive White Gaussian Noise (AWGN) model in the image denoising task, which consider the noises on the observed noisy images are randomly sampled from a Gaussian distribution and are linearly added / super-posed onto the ideal clean background. Thus, many relevant denoising algorithms / models are designed to estimate the parameters (like the standard deviation) of the Gaussian distribution Fan et al. (2019); Goyal et al. (2020); Motwani et al. (2004).

1.2.2 Representation Learning & Deep Neural Networks

Nevertheless, the earlier methods above still highly rely on human-engineered features or hypothetical models that are manually designed based on human observation, statistical understanding of the phenomena, or theories only in ideal conditions. They may not truly reflect the real-world scenarios and may probably involve human bias. Particularly for image restoration tasks, the visual degradations involved (like noise, haze, and rain) can result from complex physical processes, which may be poorly explored or even unable to fully interpret their causalities Karanam, Srinivas, and Krishna (2020); Tekalp et al. (2022); Tian et al. (2020). Therefore, with limited domain knowledge, designing general approximating models that can well describe the phenomena may be a constant gap that can never be filled.

Representation learning Bengio, Courville, and Vincent (2013); Zhong, Wang, Ling, and Dong (2016) is an advanced idea in AI that allows direct simulation of these complex processes without manually engineering features or designing hypothetical models. Rather than relying on explicit theories or prior knowledge, this idea attempts to automatically discover these features or patterns by learning from large amounts of data. Thus, this data-driven idea can be free from domain-specific knowledge and has achieved impressive results in simulating features or patterns that can hardly be described using human language or simple logic.

Deep learning Goodfellow, Bengio, and Courville (2016) is the cutting-edge technology of representation learning, which uses artificial deep neural networks (DNN) as models to learn and directly simulate the mapping from the inputs to the outputs using the back-propagation op-

timization algorithm. With considerable layers of adjustable parameters and a sufficiently large amount of data, a DNN model can theoretically approximate functions of any complexity Csáji et al. (2001); Hornik, Stinchcombe, and White (1989), which empowers simulating extremely complex features or patterns and allows learning tasks with high-level abstraction. Owning to this advanced property, deep learning methods have achieved state-of-the-art performances in a vast amount of intelligent-level tasks nowadays Dargan, Kumar, Ayyagari, and Kumar (2019); Gheisari, Wang, and Bhuiyan (2017); Pouyanfar et al. (2018).

To better simulate the complicated patterns of visual degradations such as noise, haze, and rain, modern approaches to image restoration have started to apply DNNs as the models (or the key components) in their methods. Nowadays, deep learning has become the mainstream method for many image restoration tasks and relevant applications Gui et al. (2021); S. Li et al. (2019); Tekalp et al. (2022); Tian et al. (2020); W. Yang et al. (2020a).

1.2.3 Generative Model & Generative Adversarial Networks

Unlike many other pattern recognition tasks, image restoration requires not only the modeling of visual degradations (like noise, haze, and rain) but also the study of how these degradations integrate with the background scenes to form the observed images. Most existing methods tend to apply DNN models only focusing on simulating the patterns of visual degradations for better recognition and removal, but assume these image degradations are super-posed / linearly added onto the background as a single layer Fan et al. (2019); Goyal et al. (2020); S. Li et al. (2019); Motwani et al. (2004); W. Yang et al. (2020a). Recently, a growing number of studies start to figure out that the visual degradations in image restoration tasks may not be simply added onto the background, but may have more complex compositions. Thus, many state-of-the-art methods attempt to propose various composition models in describing their integrations Hu, Fu, Zhu, and Heng (2019); Liu, Yang, Yang, and Guo (2018); McCartney (1976); Narasimhan and Nayar (2000); Tian et al. (2020); W. Yang et al. (2017) (see Chapter 2 for more detailed discussion). Noticeably, however, these composition models are still handcrafted based on hypotheses, which may not represent the true situations of integration in the real scenarios and may probably involve subjectivity or human biases. Thus, methods applying these hypothetical composition models may still not be the perfect choice for image restoration tasks, and this can also be one of the main reasons that contribute to the performance gaps of models being evaluated on these synthetic datasets than in the actual practices.

Moreover, some details and fine-grained background information of the restoration results may be lost from the observed inputs or may not be extracted, hence, additional information or extra knowledge may be required in the image restoration process to complete these missing / damaged contents. For instance, in the image dehazing task, considerable pixels areas of the background may be strictly distorted or even be completely covered by the heavy hazy / fog, which can hardly be restored by using only the information from the single input. Most existing methods tend to neglect these missing details or do not architecturally support learning about this relevant information. Some latest methods try to introduce extra DNN model(s) to learn / simulate these missing information and "add" them back in the restoration results S. Deng

et al. (2019); W. Yu, Huang, Zhang, Feng, and Xiao (2019). But these methods still tend to "deconstruct" the imaging process and, more importantly, since these extra DNN(s) are not directly optimized, they may not be well-trained and can hardly learn general knowledge for completing the missing details.

Deep-learning-based generative models (DGMs) Bond-Taylor, Leach, Long, and Willcocks (2021); Oussidi and Elhassouny (2018); Ruthotto and Haber (2021) are considered more promising models for image restoration that can cope with the above fatal problems in the existing deconstructive methods. Since DNNs and the idea of representation learning have been proven to be successful in learning the complex patterns of visual degradation, similarly, we may also apply the same idea to directly simulate / learn the complicated integration between the visual degradation and the background scenes. By considering the image restoration tasks as end-to-end generation problems (image-to-image translation), DGMs can be applied to directly simulate the entire process from the inputs to the outputs, where both the patterns of visual degradations as well as how they integrate with the background can be learned simultaneously inside the DNN models. Therefore, rather than deconstructing the complicated integration processes, these methods only learn to generate results that directly satisfy the target distribution, which does not require an understanding of the detailed mechanisms or the compositions behind. DGMs are also strong at filling up the lost information / missing pixels and generating semantically plausible restored results by transferring general knowledge learned from big data Yeh et al. (2018). In addition, generative methods also benefit from much lighter-weight models compared with those deconstructive methods using sophisticated handcrafted models and may be generally applicable to different image restoration tasks (like denoising, deraining, and dehazing) with all-in-one models, which, unlike deconstructive methods that are specific to tasks and require the specialized design of composition models for different tasks. Table 1.1 briefly summarizes the advantages of generative methods compared with the existing deconstructive methods (see Table 2.1 in Chapter 2 for more detailed comparison).

	Deconstructive methods	**Generative Methods**
simulating the real-scenarios	based on hypotheses and may involve human bias	purely data-driven and can better simulate the complicated phenomena
handling of missing details and damaged information	do not take into account or do not have an efficient approach to learn about relevant knowledge	allow completion of these missing contents during the generation process and are supported to learning about relevant knowledge from data.
complexity and scale of the models	sophisticated and large in scale	concise and lightweight
generalizability	task-specific and may require specialized designs of handcrafted models for different tasks	generally applicable for different tasks and allow all-in-one model

Table 1.1: Advantages of generative methods for image restoration tasks compared with the conventional deconstructive methods

Generative Adversarial Networks (GANs) Emami, Aliabadi, Dong, and Chinnam (2020); Goodfellow et al. (2014); Isola, Zhu, Zhou, and Efros (2017); Mirza and Osindero (2014); Radford, Metz, and Chintala (2015); Xiong, Wang, and Gao (2019); J.-Y. Zhu, Park, Isola, and

Efros (2017); J.-Y. Zhu, Zhang, et al. (2017) are one of the most promising kinds of DGMs for image restoration tasks that can generate restoration results with the highest quality among all other DGMs Bond-Taylor et al. (2021). Compared with generative methods that are based on Autoencoders (AEs) Bengio, Yao, Alain, and Vincent (2013); Bourlard and Kamp (1988); Kingma and Welling (2013); Ranzato, Poultney, Chopra, LeCun, et al. (2007), GANs introduce training loss from a discriminator network rather than directly minimizing pixel-wise consistency or optimizing latent similarities, which empowers higher visual fidelity generated results. GANs also adopt the idea of adversarial training between two networks (the generator and the discriminator), which guarantees the learning of more general features with higher-level semantics.

Therefore, recent studies Alsaiari, Rustagi, Thomas, Forbes, et al. (2019); R. Li, Cheong, and Tan (2019); R. Li, Pan, Li, and Tang (2018); Pan et al. (2020); Qian, Tan, Yang, Su, and Liu (2018); H. Zhang, Sindagi, and Patel (2019) have started to introduce GANs to their methods for image restoration tasks.

1.3 Problem Statement

1.3.1 Quantitative Performance of Existing Generative Methods

Nevertheless, despite the promising prospects of DGMs (particularly GANs) for image restoration tasks, current methods that apply these models to image restoration still have plenty of room for improvement in performance compared with those conventional methods using handcrafted composition models. With considerable advantages mentioned earlier, methods using the generative idea are supposed to perform better on image restoration tasks than those conventional deconstructive approaches. However, existing methods applying these generative models have not yet achieved competitive performance compared with the state-of-the-art results using handcrafted composition models. Notably, most DGMs, especially GAN models for the conventional generation problems of image-to-image translation can be directly used in image restoration tasks and can obtain pretty nice-looking restoration results. Whereas, in terms of their quantitative performance, these methods still tend to have a large gap to fill in evaluating their restoration results on the benchmarks, or may require much larger amounts of training data so as to reach a similar performance. The reasons behind this poor quantitative performance of the existing generative methods can be worth exploring.

1.3.2 Black Boxes of Deep Learning

Deep neural networks have long been viewed as black boxes due to lacking interpretability Goodfellow et al. (2016). Especially for DGMs, the generation processes are fully conducted by the parameters inside the generator networks, where the internal operations and the reliability of the generated results are still unclear.

Unlike the deconstructive methods that have been well-studied and whose mechanism can be easily and intuitively explained, generative methods are recently introduced to the image

restoration tasks, which tend to be under-exploration and require a deeper understanding for us to validate their potential and analyze the designs of relevant network models.

In the past decades, considerable theories were proposed trying to open the black box of deep learning Buhrmester, Münch, and Arens (2021), but only limited studies have attempted to explain the DGMs, and there is still no perfect theory to fully understand what happens inside these DGMs. Thus, nowadays, the interpretation of deep neural networks, especially DGMs, remains an open challenge.

Therefore, a theoretical explanation of their learning processes and related mechanisms can be of urgent need. It can be helpful for understanding and solving the problems of existing generative methods that lead to poor quantitative performance on image restoration tasks. Moreover, the relevant theory may also attach great significance to the analysis and design of network models, prediction of the learning behaviors as well as the enhancement of performances on various tasks and applications.

1.4 This Thesis

1.4.1 Issues of Applying Conventional GANs to Image Restoration

Regarding the first problem, in this thesis, we consider that the general DGMs designed for conventional generation tasks may not be directly applicable for image restoration. We noticed that: existing generative methods for image restoration tend to be direct applications of the conventional DGMs, which, however, were originally designed for more general generation tasks that vary much from the tasks in image restoration. Thus, these conventional models may not be suitable for image restoration tasks, where some key issues may be left unsolved, contributing to their performance gaps.

Therefore, to understand the reasons behind the poor quantitative performance of the existing generative methods, we need to explore the issues when directly applying conventional DGMs to image restoration tasks.

In this thesis, we identified three key issues of applying conventional GAN models to image restoration tasks, which include (1) over-invested abstraction process, (2) inherent details loss, as well as (3) vanishing gradient and imbalance training. We theoretically analyzed and explained these issues and proved their existence in current GAN-based image restoration methods and have significant impacts on their models' performances with empirical evidence.

1.4.2 Information Bottleneck in DGMs

Information Bottleneck (IB) principle Tishby, Pereira, and Bialek (2000); Tishby and Zaslavsky (2015) is currently one of the most sounded theories proposed to understand the deep learning processes. It tries to open the black boxes of deep neural networks by employing the information-theoretic method, which helps to quantify the information flow in a general deep neural network model by estimating the mutual information between different layers of inputs and outputs. This theory interpreted the process of deep learning as an information trade-off

between compression and prediction, which helps to understand the reasons behind the remarkable generalization performance of deep neural networks. This theory is also well consistent with the learning behaviors of neural networks and can properly explain many experimental phenomena we observe. However, the principle is only proposed to explain the information compression and extraction process, which is more suitable for the discriminative models in deep learning rather than DGMs.

In this study, we deem this IB theory can be helpful for us to understand the mechanisms behind DGMs, and may even help to explain the reasons behind the performance gaps of existing generative methods for image restoration tasks. By using this theory, we can analyze the information flow inside the deep neural network models, which helps the design of relevant network structures for different tasks. Moreover, by exploring the optimization objectives of the network models, we can enhance existing loss functions to improve the training process or customize training strategies for specific tasks.

In this study, we propose an information-theoretic framework based on the information bottleneck principle to analyze DGMs as well as their learning process for the image restoration tasks. Based on the theory, we analyze the information flow in GAN-based generative methods for image restoration and identify three sources of information inside the restoration processes. Moreover, we calculate the optimal amount of information for each of these sources, which reveals the optimization boundaries for different parts of the models and can guide the analysis and design of network models for image restoration and even many other tasks.

1.4.3 Summary of Problems & Hypotheses

To sum up, two critical problems are our main focus in this thesis:

- Despite their promising prospects, existing generative methods fail to achieve satisfactory performance on image restoration compared with the conventional deconstructive methods.
- The mechanisms behind deep generative models are rarely explored and there is no mature theory available for the analysis of these generative methods on image restoration tasks.

Regarding the problems above, we have the following two hypotheses respectively:

- Existing generative methods tend to be direct application of the conventional DGMs, which may not be suitable for the image restoration tasks and have various problems left unsolved.
- we can extend the IB Principle to construct a theoretic framework for explaining the DGMs and thus help the analysis of generative methods and instruct the design of network models for image restoration as well as other tasks.

1.4.4 Overview

The remaining parts of this thesis will be organized as follow:

In Chapter 2, we reviewed the existing methods, particularly deep-learning-based methods for the image restoration tasks including image denoising, image dehazing, and deraining. We divided existing methods into two categories: deconstructive methods and generative methods and demonstrate their progress respectively.

In Chapter 3, we theoretically analyzed the flow of information in these generative models and proposed our information-theoretic framework based on the information bottleneck principle. We also indicate that information involved in the generated results originates from three sources: (i) high-level information extracted by the encoder network, (ii) low-level information from the source inputs that retained or pass directed through the skip connections, and, (iii) external information introduced by the learned parameters of the decoder network during the generation process. In the tasks of image restoration, we consider this information mainly corresponds to (i) image degradations to be removed, (ii) details and background information, and, (iii) information that is lost or absent from the inputs to be recovered in the generated outputs, respectively.

Based on this analysis, in Chapter 4, we identified three key issues that probably hinder the models' performances when directly adopting GANs to the task of image restoration: (i) models designed for conventional generation tasks involve over-invested abstraction processes that may not be helpful for the learning of features of image degradations; (ii) network structure of the generator models tend to inherently discard details information which can be more fatal in the image restoration task; (iii) existing measures for evaluating generated results fail to provide efficient gradients for further optimization to learn external knowledge for generative contents. We formulate these problems by providing corresponding definitions and theoretical explanations in this chapter.

In response to these problems, we proposed corresponding solutions and suggestions to improve GANs models' performances on relevant tasks in Chapter 5, including optimization of network structure, enhancing details extraction, accumulation, and retention, as well as alternating the measures of the loss function.

In Chapter 6, we proved the existence of the identified issues above with empirical evidence respectively and examined the proposed methods with experiments.

Ultimately, the thesis ends with a conclusion (Chapter 7), summarizing the entire thesis and discussing the limitations as well as the future works of this research.

Chapter 2

Literature Review

Based on the modeling of how image degradations (like noise, haze, and rain) integrate with the background scenes to obtain the observed images, methods for image restoration can be categorized into two groups: deconstructive methods and generative methods. Deconstructive methods try to describe the above integration by handcrafting physical models, while generative methods tend to learn these integrations in an end-to-end data-driven manner.

In this chapter, we are going to review the progress of different image restoration methods following the two categories above, introduce their key ideas and identify their existing problems of concern.

2.1 Deconstructive Methods for Image Restoration

2.1.1 Linearly Additive Composition

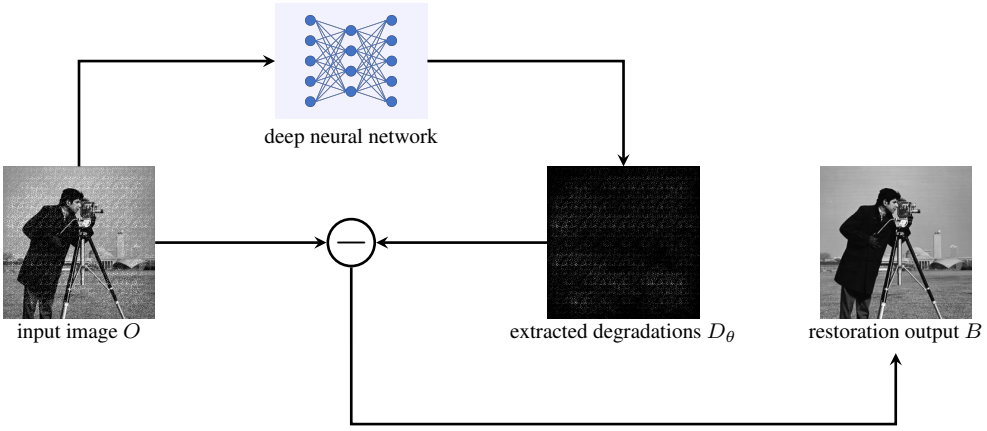

Figure 2.1: Schematic diagram of deep-learning-based deconstructive methods using the linearly additive composition model for image restoration tasks

Early studies of image restoration assume that the visual degradations are single layers that are super-posed / linearly added onto the background scenes, and relevant methods mainly focus on modeling the patterns of these degradations in order to better recognize and remove them.

Initially, people tend to develop various hypotheses and design relevant models to describe their features and patterns. For example, the early idea of image denoising problem considers noise is randomly sampled from a Gaussian distribution, which is generally known as Additive

White Gaussian Noise (AWGN). Thus, these methods focus on finding the parameter settings of this Gaussian distribution, so as to simulate and separate noise from the input image. Later literature extended the modeling of noise to other types of distributions, including quantisation noise, impulse noise (salt and pepper), Poisson noise, and speckle noise et al Fan et al. (2019); Goyal et al. (2020); Motwani et al. (2004). Whereas, the most recent studies consider the formulation of noise in real noisy images can be a hybrid of different distributions, or even blind to us for designing relevant models Tian et al. (2020). In addition, for more complicated image degradations like haze and rain, people designed models such as sparse coding L.-J. Deng, Huang, Zhao, and Jiang (2018); Kang, Lin, and Fu (2011); Luo, Xu, and Ji (2015); L. Zhu, Fu, Lischinski, and Heng (2017) and Gaussian Mixture Model Y. Li, Tan, Guo, Lu, and Brown (2016) to describe the characteristics and simulate the patterns of haze and rain. However, these models are designed manually based on human observation or statistical analysis with limited data, which may not accurately reflect the rain pattern in real situations.

Later studies start to apply deep learning methods to automatically learn these features of visual degradations from large amounts of data. Zhou et al. 1987 is the first attempt to use a neural network as the model for image restoration on the image denoising task. After which, Jain and Seung 2008 first proposed to use Convolutional Neural Network (CNN) for this task. But the neural networks here are shallow and different from the ones used in modern deep learning technology. It is after 2012 when deep learning had achieved great success in a variety of Computer Vision tasks, people began to explore using "deeper" neural networks for image denoising K. Zhang, Zuo, Chen, Meng, and Zhang (2017), as well as applying this ground-breaking technology widely to other image restoration tasks and applications. Cai et al. 2016 first introduce deep learning technology to the task of image dehazing, and subsequently, this idea is also adopted to image deraining tasks by Yang et al. 2017. The representation learning of deep neural networks gets rid of human biasing and achieves significant improvement in performance. Nowadays, deep learning is becoming the mainstream method for simulating the various patterns of visual degradations in image restoration tasks.

Nevertheless, for the question of how these patterns of degradations integrate with the background scenes in the observed images, most existing methods are still based on the Additive Composite Model, which simply assumes a layer of these degradations is linearly added onto the backgrounds:

$$O = B + D \tag{2.1}$$

where O and B denote the observed degraded image and the targeted clean background layer to be restored respectively, and D here represents the layer of image degradations simulated by handcrafted or deep-learning-based models.

2.1.2 Handcrafted Composition Models

A growing number of studies started to figured out that these image degradations may consist of multiple layers, and may not be simply super-posed onto the background image.

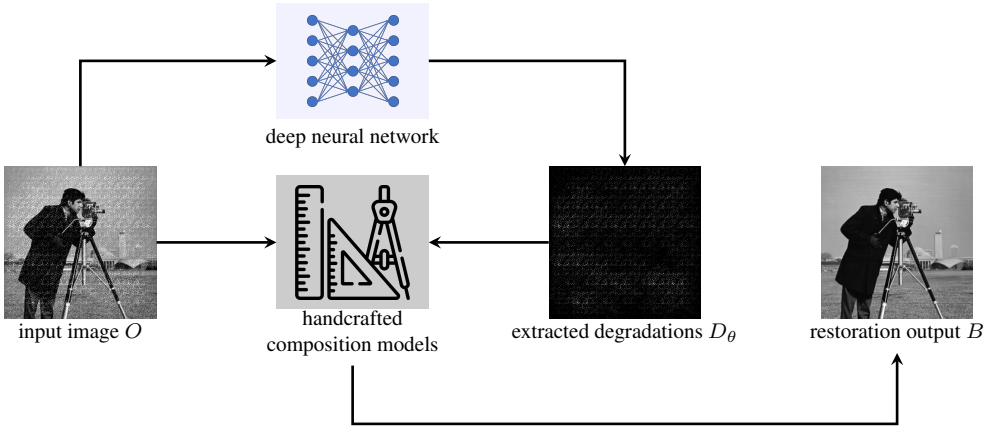

Figure 2.2: Schematic diagram of deep-learning-based deconstructive methods using handcrafted composition models for image restoration tasks

In the image deraining task, Yang et al. 2017 points out that rain in the observed images consists of not only a single layer, but accumulations of multiple rain streak layers. Thus, they proposed a heavy rain model to describe how multiple layers of rain accumulation blend in with the background scene, where the attenuation of atmospheric transmission caused by ubiquitous turbid media is also taken into account.

$$O = \alpha \left(B + \sum_{t=1}^{T} R_t \right) + (1 - \alpha) A \tag{2.2}$$

R_t here denotes a single rain layer and t indexes the layer with T as the total number of layers. Referring to the imaging models for fog and haze, the global atmospheric light A is introduced to represent the physical atmospheric scattering, and α is the atmospheric transmission.
As an extension of this model, Liu et al. 2018 further divides rain into transparent and opaque rain (rain streaks that completely covered the background scenes) and proposed their occlusion-aware model. Based on statistical analyses on its physical properties, Hu et al. 2019 discovered that the appearance of rain is related to the scene depth: rain appears as rain streaks when it is close to the camera, while it tends to be fog-like rain accumulations when it is far away. Based on this observation, they proposed a depth-aware rain model by considering image scene depths.

$$O = (1 - R_s - F) B + R_s + AF \tag{2.3}$$

$$R_s = \sum_{t=1}^{T} R_t \odot e^{-\alpha max(d_m, d)} \tag{2.4}$$

$$F = 1 - e^{-\beta d} \tag{2.5}$$

where R_s and F denote the rain streak layer and fog (rain accumulation) layer respectively. d is the scene depth meaning that the distance between the object and the camera, and d_m is a

threshold defined to represent the maximum distance that rain streak can be observed (farther will be fog). α and β are attenuation coefficients that control the rain streak intensity and the thickness of fog respectively.

Some other methods S. Deng et al. (2019); W. Yu et al. (2019) considered that details in the background may be covered or seriously distorted by the rain layer. Thus, these methods attempted to learn these lost details, and try to add them as a supplementary in the resulted image S. Deng et al. (2019), or conduct a coarse-to-fine process, where the lost details can be recovered at the fine-grained stage W. Yu et al. (2019).

$$O = B - D + R \tag{2.6}$$

where D denotes the lost details in the background.

Whereas in the image dehazing task, the deconstructive idea had been proposed early in the causality analysis stage of studies. The fundamental *atmospheric scattering model* was proposed by McCartney 1976, after which Narasimhan and Nayar 2000 complement this model with theoretical description.

$$O = TB + (1 - T)A \tag{2.7}$$

$$T = e^{-\beta d} \tag{2.8}$$

similarly, A and d denote the global atmospheric light and scene depth respectively, while T here represent the matrix of transmission, with β as the scattering coefficient of the atmosphere.

To estimate the transmission matrix T above, He et al. 2010 proposed the famous algorithm of Dark Channel Prior

$$T = 1 - w \min_{c}(\min_{y}(\frac{O^c(y)}{A^c})) \tag{2.9}$$

where c represents the 3 channels of red, green blue (RGB) of the image, y represents an element of pixel in the image and w is a constant of coefficient.

With the outstanding performance of deep learning technique, current methods try to learn this transmission matrix T using deep neural networks.

2.1.3 General Review on Deconstructive Methods

In general, all these methods attempt to deconstruct the imaging process or physics of integrations using human-designed composition models, which:

- assume the visual degradations and the background scenes are distinctly disjoint, which may not truly reflect how they actually integrate under the real-world situation;
- highly rely on human observations, hypothesis, and domain priors, which may involve human bias;

- architectures built based on these handcrafted models are often complicated and large in scale, which may easily result in over-fitting;
- missing details and damaged information are ignored, or do not have efficient approaches to support the learning and involving of relevant knowledge to complete these parts of the information;
- task-specific that may require specialized designs of composition models for different image restoration tasks.

Since the data-driven methods of deep learning have achieved great success in simulating the patterns of these image degradations, the same idea of representation learning may also be more promising in modeling the integration of these patterns with the background than those handcrafted models.

2.2 Generative Methods for Image Restoration

As a solution, generative methods allow approximating the entire image restoration process as an end-to-end mapping from inputs directly to the targeted outputs, where both the patterns of visual degradations and their integration with the background can be learned simultaneously inside the DGMs. Therefore, handcrafting composition models are no longer needed. Different from the deconstructive idea above, generative methods learn to simulate the target distribution directly and attempt to generate data that looks like these targets. Rather than understanding the mechanism of how these visual degradations blend in with the background, these methods care more about the generated results. Moreover, compared with those deconstructive methods, models in generative methods are intrinsically more concise and smaller in scale, making them less likely to overfit.

2.2.1 Generative Methods based on Autoencoders

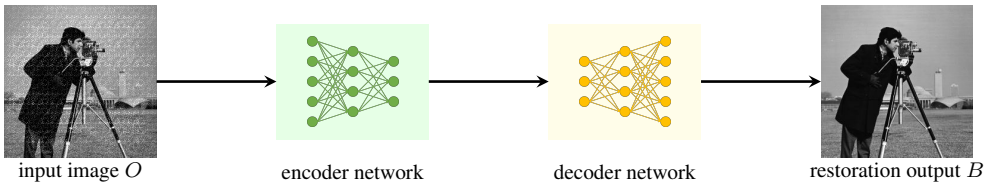

Figure 2.3: Schematic diagram of Autoencoders-based generative methods for image restoration tasks

The simplest form of generative methods is based on Autoencoders Bourlard and Kamp (1988); Kingma and Welling (2013); Ranzato et al. (2007), where the generator models learn by minimizing pixel-level similarities between output and the target ground truth. Autoencoders (AEs) are the most commonly-used generative network models in deep learning. As the extensions of Restricted Boltzmann Machines (RBM) Hinton (2002); Hinton and Salakhutdinov

(2006); Hinton, Sejnowski, et al. (1986); Salakhutdinov, Mnih, and Hinton (2007); Smolensky (1986), they are originally designed for the unsupervised learning of data compression or extraction of lower-dimensional representations from the high dimensional inputs Bourlard and Kamp (1988). More specifically, these network models are trained to perform data reconstruction: given a high-dimensional data as the input, the encoder part of the network tries to extract a much lower-dimensional variable as its latent embedding, and then using only the limited information of this embedding, the decoder part of the network tries to generate (reconstruct) an output as similar to the source input as possible. In this way, we can extract a meaningful representation (or compression) of the data. By simply replacing the output targets with the generated results, these models can therefore be generalized as generative models for various generation tasks. Noticeably, the measures of similarity between the generated outputs and the corresponding targets (or inputs for reconstruction training) are often calculating their differences in pixels' values. Therefore, the generated results are expected to be identical to their targets without any variation allowed.

Autoencoders have been applied to the image denoising task for a long time, and studies have shown that they are better than many traditional methods in handling blind denoising, that is, noise with unknown patterns Majumdar (2018). Nevertheless, Autoencoders are still not the mainstream methods for image denoising, and some studies point out that these autoencoder-based generative methods may introduce blurring and may reduce the verisimilitude in the generated results Zhao (2012).

Remark. The concept of Denoising Autoencoder (DAE) is different from applying autoencoders to image denoising tasks. In DAE, the models try to add extra noise to their inputs so as to enhance the robustness and generalizability of the learned representations.

Chen and Lai 2019 first apply Autoencoder to the task of image dehazing, but the performances of this model are less competitive: results on some datasets are far less than the corresponding results of DehazeNet Cai et al. (2016), which is the first methods to use deep learning for image dehazing mentioned earlier. Bennur and Gaggar 2020 also proposed an image dehazing model using an Autoencoder, whereas, instead of comparing their results with others, this work emphasizes the advantage of their method lies in a lighter-weight model compared to other approaches.

For the image deraining task, Du et al. 2020; 2020 adopt conditional variational autoencoders (c-VAEs) Kingma and Welling (2013) as their deraining models. However, in order to make their restored images look more plausible, this method generates a massive amount of predictions at a time and then applies their weighted average as the final dehazed results, which can be computationally costly and the average values of pixels may be distorted. Fu et al. 2019 also applies autoencoder-structure as its generator network in a bid to be conspicuous as a lighter-weight model. However, domain-specific knowledge of Gaussian Laplacian image decomposition is used as the down-sampling process, which may not be regarded as a purely end-to-end learning process.

In summary, we can find that: Autoencoder trained simply based on pixels-wise similarity may not guarantee the generation of high-quality results. And for the image restoration

tasks, the poorly trained Autoencoder models may introduce extra noises or blurring, leading to distortion in the restoration results.

2.2.2 GAN-based Generative Methods

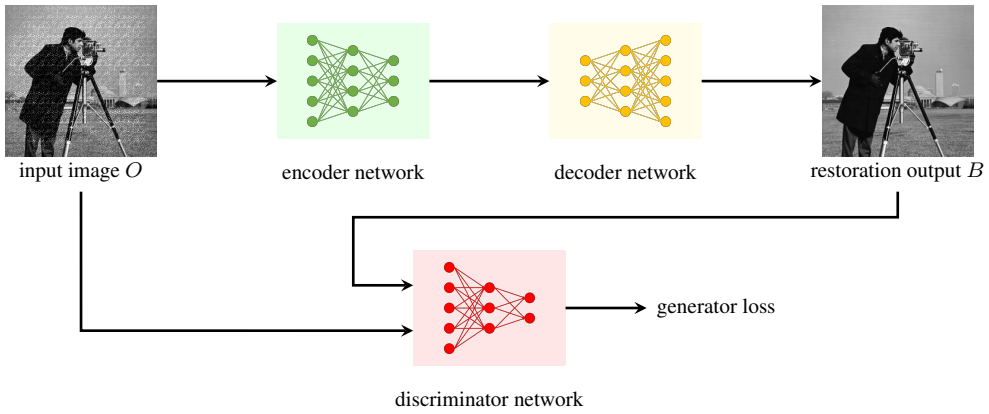

Figure 2.4: Schematic diagram of GAN-based generative methods for image restoration tasks

Generative Adversarial Networks (GANs) Emami et al. (2020); Goodfellow et al. (2014); Isola et al. (2017); Mirza and Osindero (2014); Radford et al. (2015); Xiong et al. (2019); J.-Y. Zhu, Park, et al. (2017); J.-Y. Zhu, Zhang, et al. (2017) are considered more promising generative methods for image restoration tasks. In addition to a generator network, GANs introduce a discriminator network with an adversarial training strategy to provide guidance for optimizing its generated outputs. Therefore, rather than to identically approximate the generation targets by directly minimizing pixel-wise similarity or optimizing latent consistency in the generative methods based on Autoencoders, GANs allow the generated results only to share consistency with the targets in the higher-level semantics. This empowers the generation of more eidetic results and brings in impressive performances on various generation tasks. For image restoration tasks, GANs can be more advanced compared with the Autoencoders-based generative methods in generating results with higher visual fidelity and restoring missing information with more plausible contents and details, compared with Autoencoders-based generative methods.

In fact, most typical GAN models for image-to-image translation (like pixel2pixel Isola et al. (2017) and CycleGAN J.-Y. Zhu, Park, et al. (2017)) can be directly applied to the tasks of image restoration and obtain plausible restoration results. However, the quantitative evaluations of their performances may not be satisfactory compared with the traditional deconstructive methods using complicated handcrafted composition models.

Considerable studies have attempted to apply GANs to perform image denoising, but many of them only demonstrate their visual results without quantitative comparisons of their models' de-noising performance Alsaiari et al. (2019). Some of the works emphasize their advances in blind denoising J. Chen, Chen, Chao, and Yang (2018), some indicate their contributions in the training of models using unpaired data Hong, Fan, Jiang, and Feng (2020), some others focus on the image denoising problems on domain-specific data and applications such as medical images Kim and Lee (2020); Q. Yang et al. (2018), bio-statistics data (cell images) S. Chen, Shi, Sadiq,

and Cheng (2020), and digital radiography (X-ray images) Sun, Liu, Cong, Li, and Zhao (2018) et al.

The more improvements GANs provide are seen in the more complicated image restoration tasks like image dehazing. AOD-Net. B. Li, Peng, Wang, Xu, and Feng (2017) proposed an all-in-one GANs for image dehazing by designing the generation based on the fore-mentioned atmospheric scattering model McCartney (1976); Narasimhan and Nayar (2000), but obviously, this method still relies on the handcrafted composition model which may not be regarded as an end-to-end generation process. Li et al. 2018 directly apply a Conditional GAN (CGAN) Mirza and Osindero (2014) model to the dehazing tasks and have achieved impressive performances. Nevertheless, the model applies a weighted sum of multiple loss functions as its training objectives, which introduce considerable hyper-parameters to adjust. Moreover, the use of perceptual loss requires a feature extractor that is well pre-trained on other datasets, which involves external knowledge from extra datasets and can be unfair to compare with other methods trained using limited data. Similarly, DHSGAN method Malav, Kim, Sahoo, and Pandey (2018) also uses an end-to-end GAN model as their dehazing architecture, but this method initializes the parameters of their model by pre-training on other datasets, which is also likely to include external information more than those contained in the training datasets. Engin, Genç, and Kemal 2018 also explored the possibility of training dehazing models in an unsupervised manner without using paired data, they directly apply a CycleGAN J.-Y. Zhu, Park, et al. (2017) as their dehazing model and demonstrate promising results in the real-world practices. Since no ground truth is used in this scenario, we have no measure to quantitatively evaluate the dehazing performances of their model.

ID-CGAN H. Zhang et al. (2019) is the first method that applies GANs to the image deraining task. It adopts the pixel2pixel Isola et al. (2017) (a typical architecture of CGAN for image-to-image translation), introduces DenseNet Huang, Liu, Van Der Maaten, and Weinberger (2017) structure in each up-and-down-sampling layer of the UNet Ronneberger, Fischer, and Brox (2015) generator (which similar to the Tiramisu Network Jégou, Drozdzal, Vazquez, Romero, and Bengio (2017)), and propose a multi-scale discriminator. However, the proposed modifications do not significantly improve in performance compared with the original pixel2pixel GAN, and we noticed that serious details loss exists in the generated results. Qian et al. 2018 introduced the attention mechanism to GANs and proposed an attentive-recurrent network to extract an attention map of the rain-blurred regions and pass it together with the observed image to the generator model. Nevertheless, the entire model is optimized in multi-paths, where the resulted performances can be sensitive to the quality of the extracted attention map as well as the coefficient between weighting items of the loss functions. Moreover, this model focus only on the removal of adherent raindrops on the glasses, which is apart from the main-stream deraining tasks that try to remove rain streak and rain accumulations. Li et al. 2019 also involved a GAN model in their deraining architecture, but it is only used at the end of their model for refinement and the overall architecture is still largely based on a handcrafted rain-background composition model. Therefore, the model is in fact not an end-to-end generation model. Moreover, the training of this model requires extract supervision other than the

input and ground truths, which may largely depend on the dataset and may hardly generalize for real-world deraining practices.

2.2.3 General Review on Generative Methods

As summarized in Table 2.1, compared with deconstructive methods that are based on hypothetical composition models, generative methods, especially GAN-based generative methods, have shown advantages in a variety of aspects. Thus, generally speaking, these methods are supposed to achieve better performance in image restoration tasks.

<table>
<tr><th colspan="2">Deconstructive Methods</th><th colspan="2">Generative Methods</th></tr>
<tr><th>Linearly Additive Composition</th><th>Handcrafted Composition</th><th>Autoencoder-based</th><th>GAN-based</th></tr>
<tr><td>are only hypothetical and may not truly reflect the real-world scenarios;</td><td>can better simulate the real scenarios but still tend to be restricted by ideal assumptions;</td><td colspan="2">can perfectly approximate real-world situations as long as the amount of training data is enough.</td></tr>
<tr><td colspan="2">may involve human bias;</td><td colspan="2">purely data-driven thus can avoid human bias.</td></tr>
<tr><td>reasonable in models' complexity and scale;</td><td>models are often sophisticated and of massive scale;</td><td colspan="2">models tend to be concise and lightweight.</td></tr>
<tr><td>inference can be fast;</td><td>inference speed tends to be slow;</td><td colspan="2">inference can be fast.</td></tr>
<tr><td>training is conducted in multiple paths and thus the parameters inside are not directly optimized;</td><td>parameters may not be directly trained / optimized, or may even require multi-stages of training;</td><td colspan="2">allow direct optimization and efficient updating of parameters.</td></tr>
<tr><td>do not support completing missing / damaged details and the relevant information is ignored;</td><td>only a few methods allow the completion of missing / damaged details but do not have an efficient way to learn relevant information and knowledge;</td><td>functionally fully support the completion of missing / damaged information but still do not have an efficient approach to guarantee the learning of relevant knowledge in general;</td><td>are fully capable of completing missing / damaged information and have corresponding optimization objectives to ensure the learning of relevant knowledge as well.</td></tr>
<tr><td>use similar models for all different image restoration tasks</td><td>tend to be task-specific and may require specialized designs of composition models for different task;</td><td colspan="2">can be generally applicable to different image restoration tasks and allow all-in-one model.</td></tr>
</table>

Table 2.1: Comparison among different kinds of methods for image restoration

Nevertheless, we found that exiting GAN-based methods for image restoration tend to be direct applications of the conventional GAN model. Despite the prominence of GANs in simulating the patterns of different kinds of visual degradations as well as their physical integrations with the background scenes under real-world scenarios, these GAN models are not originally designed for tasks of image restoration, and thus may not be directly applicable. By reviewing the relevant existing methods we can notice that: these conventional GANs may suffer from problems like distortion, details loss, and insufficient accuracy in their generated results, contributing to poor quantitative performances on image restoration tasks compared with the traditional deconstructive methods.

Chapter 3

Information-theoretic Frameworks

Information Bottleneck (IB) Principle Tishby et al. (2000); Tishby and Zaslavsky (2015) provides an approach for us to understand the learning process inside the black boxes of deep neural networks. Employing the information theory, it quantifies the information flow in the network models, generally explains the learning dynamics, and reveals the generalization-ability of the deep learning methods. By extending this theory, we may be able to further explain the relevant mechanisms behind deep generative models, and therefore, assist in the analysis of generative methods for image restoration tasks.

In this chapter, we will first introduce the information bottleneck (IB) principle and relevant basics in information theory. Based on this theory, we will try to explain the learning processes of three standard deep generative models in GANs as well as the conventional idea of the generative methods for image restoration. Then, we will propose our new interpretation of the generative models applied to the image restoration tasks. Based on this theoretical framework, we analyzed the information flow in deep generative models for image restoration tasks and identified three sources of information required for the generation of restoration results.

3.1 Information Bottleneck Principle

Before going deep into the IB theory, we may first examine some theoretical information quantities upon which the theory is based.

3.1.1 Information Theory

Information theory is invented by Shannon Shannon (2001) to determine the number of bits needed to transmit a message over a noisy channel. It provides the basic idea to measure the influence between variables, and quantify the abstract concept of "information".

Definition 3.1.1 (Entropy)**.** Given a random variable $X \in \mathcal{X}$ (where $\mathcal{X}$ denotes its support) and the probability distribution of its realizations $p(x)$. Then the entropy of $X \sim p(x)$ is defined as the uncertainty involved with its distribution:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log(p(x)) \tag{3.1}$$

Entropy $H(X)$ reflects the minimum description length needed to determine the exact value

of the random variable X. The value of entropy reaches its maximum when the distribution is uniform, and will be zero if it is deterministic. Thus, in general, entropy formally quantifies the amount of information in X.

Remark. The units of entropy accord with the base of the logarithmic function applied, which, in convention, is in bits and hence a base of 2 is used.

Definition 3.1.2 (Mutual Information)**.** Consider $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ as random variables with the joint distribution function of $p(x, y)$, where $(X, Y) \sim p(x, y)$. Let $p(x)$ and $p(y)$ be their marginal distributions respectively. The mutual information between X and Y is defined as:

$$I(X;Y) = \sum_{x \in \mathcal{X}, Y \in \mathcal{Y}} p(x,y) \log(\frac{p(x,y)}{p(x)p(y)}) \tag{3.2}$$

Mutual information $I(X;Y)$ measures the statistical dependence between variables X and Y, which can be interpreted as the reduction of uncertainty regarding one variable when the information of the other variable is known.

Theorem 3.1.1.

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) \tag{3.3}$$

Therefore, mutual information $I(X;Y)$ can be regarded as the amount of information shared between X and Y. If X is completely known given Y, i.e. $H(X \mid Y) = 0$, then their mutual information $I(X;Y) = H(X)$. $I(X;Y) = 0$ if and only if X and Y are independent.

Based on the theorem above, we can also obtain another essential property about mutual information, which is the data processing inequality (DPI):

Theorem 3.1.2 (Data Processing Inequality (DPI)). Suppose the random variables X, Y and Z form a Markov chain $X \to Y \to Z$ (which is defined as $p(x, y, z) = p(x, y)p(z \mid y)$), then we will have

$$I(X;Y) \geq I(X;Z) \tag{3.4}$$

DPI implies the information of X contained in Y will only decrease or keep unchanged if we apply another processing on Y, which is an important concept in describing information transmission, compression, and extraction processes.

3.1.2 Information Bottleneck Principle for Deep Neural Networks

The information bottleneck principle is propose to analyze the process of information extraction Tishby et al. (2000), which was later introduced to explain the discrimination process (feature extraction) inside the deep neural network models Tishby and Zaslavsky (2015).

Given a deep neural network model that tries to extract information from a random variable input X that is relevant to another random variable Y, its latent output $\tilde{X}$ from any hidden layer inside the network form a Markov Chain:

$$Y \to X \to \tilde{X} \tag{3.5}$$

The principle reveals that the learning processes of the deep neural network models are to optimize:

$$\min[I(X;\tilde{Y}) - \beta I(Y;\tilde{Y})] \tag{3.6}$$

where β is a positive Lagrange multiplier that trades-off between the two terms.

In other words, the optimization objectives of the deep neural networks are to minimize the mutual information between the extracted representation $\tilde{X}$ and the input X ($I(X;\tilde{X})$) while at the same time, maximize the mutual information between $\tilde{X}$ and the target output Y ($I(Y;\tilde{X})$). Therefore, the goal of deep learning can be interpreted as an information-theoretic trade-off between compression and prediction.

3.2 Information-theoretic Frameworks of Deep Generative Models

Since the above idea is only proposed to explain the information compression processes (i.e. the extraction of features and representations) inside the deep neural networks, it tends to be more applicable for models doing tasks like classification or prediction, and may not be used to interpret the deep generative models. Referring to later works using this theory Alemi, Fischer, Dillon, and Murphy (2016); X. Chen et al. (2016); Jeon, Lee, Pyeon, and Kim (2021), we extend and generalize the IB theory specifically to different deep generative models (See Fig. 3.2, 3.4 and 3.6).

3.2.1 Information Dynamic in the Original GAN Model

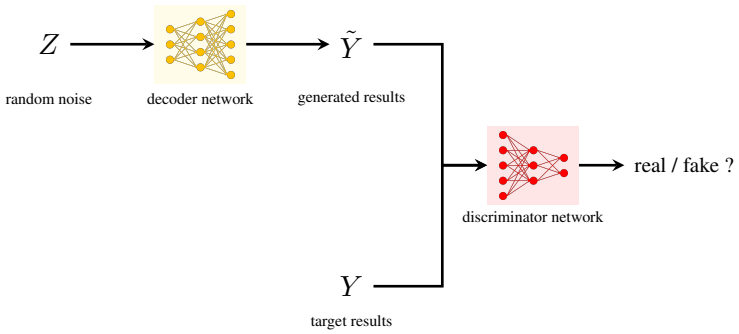

Figure 3.1: Schematic diagram of the original GAN model

The original GAN model Goodfellow et al. (2014) (See Fig. 3.1) is proposed only to generate a specific kind of data, such as generating images of different human faces. Thus, it can be regarded as a decoder network with random noise Z as the inputs and learns to generate results $\tilde{Y}$ that approximate the targets Y guided by the discriminator (See Fig. 3.2, where indirect information flow, i.e. information retrieved through the guidance of the discriminator, is represented by dotted arrows). In line with the information bottleneck for Variational Autoencoder (VAE) Kingma and Welling (2013) proposed by Alemi et al. Alemi et al. (2016), we can

induct that the optimization objective of this basic GAN model is to reach a balance between the information from its inputs Z and the targets Y in the generated results $\tilde{Y}$:

$$\max[\alpha I(Y;\tilde{Y}) + (1-\alpha)I(Z;\tilde{Y})] \tag{3.7}$$

where α here is a coefficient to control the balance between the two terms.

Noticeably, the random noise Z here is responsible for adding variations (mainly low-level details) to the generated results (such as controlling the generation of different faces) and thus is independent of Y ($I(Y;Z) = 0$).

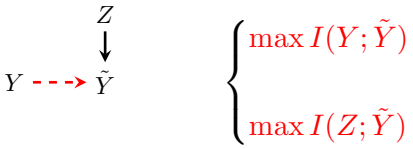

Figure 3.2: Information flow and optimization Objective of the original GAN model

3.2.2 Information Dynamic in the Conditional GAN Model

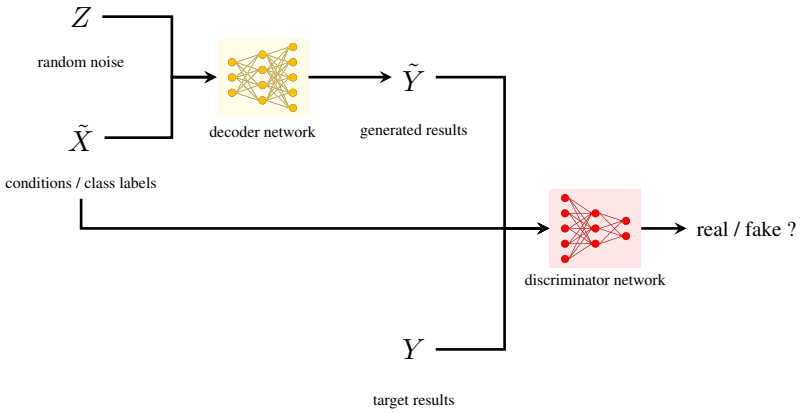

Figure 3.3: Schematic diagram of the Conditional GAN (CGAN) model

Apart from the random noise Z, Conditional GAN (CGAN) Mirza and Osindero (2014) (See Fig. 3.3) and InfoGAN X. Chen et al. (2016) take extra inputs of conditions / class labels (here denoted as $\tilde{X}$) and expect the information of $\tilde{X}$ can appear in / can be fully extracted from their corresponding generation results $\tilde{Y}$. Thus, in these cases, this informative part of inputs $\tilde{X}$ is supposed to be highly related to the targets Y, where Y often contain all information of $\tilde{X}$ ($I(Y;\tilde{X}) = H(\tilde{X})$). Therefore, information in Y is guiding the generation of $\tilde{Y}$ in two paths: $I(Y;\tilde{X};\tilde{Y})$ and $I(Y \mid \tilde{X};\tilde{Y})$ respectively (Fig. 3.4).

Similar to above, these kinds of GAN models optimize the corresponding information as follow:

$$\max[\alpha I(Y;\tilde{Y}) + \beta I(Z;\tilde{Y}) + (1-\alpha-\beta)I(\tilde{X};\tilde{Y})] \tag{3.8}$$

where α and β are corresponding weight coefficients.

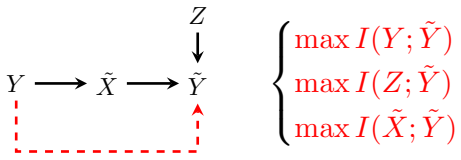

Figure 3.4: Information flow and optimization objective of Conditional GAN (CGAN)

3.2.3 Information Dynamic in the GAN Models for Image-to-image Translation

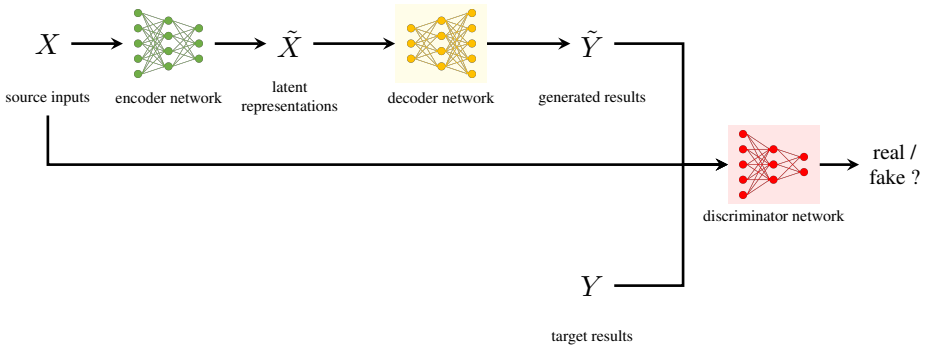

Figure 3.5: Schematic diagram of the GAN models for image-to-image translation

Models above only play the role of generation, where inputs to the networks (decoder) are already highly condensed and often in low-dimensionality. Whereas for tasks like image-to-image translation Isola et al. (2017); J.-Y. Zhu, Park, et al. (2017), the inputs to the models (such as images) are of high-dimensionality and may involve large amounts of irrelevant information. Therefore, encoder networks are equipped in these kinds of GAN models for extracting decisive features $\tilde{X}$ from the high-dimensional inputs X before passing them to the decoders for the generation process (Fig. 3.6). The overall training objectives of the models, therefore, consist of components for both the encoder network (formula in blue: compressing information and fitting the targets' features) and the decoder network (formula in red: optimizing generation results).

$$\max[\alpha I(Y;\tilde{Y}) + \beta I(\tilde{X};\tilde{Y}) + \gamma I(Y;\tilde{X}) + (1-\alpha-\beta-\gamma) I(\tilde{X};\tilde{X})] \tag{3.9}$$

where α, β and γ are corresponding weight coefficients.

Notably, the features $\tilde{X}$ to be extracted are supposed to be the information in the inputs X that can be utilized and help the generation and simulation of the targets Y (i.e. information shared between X and Y: $I(Y;X)$). In the conventional image-to-image translation problems, this allows the generated results $\tilde{Y}$ to share consistency with their input images X only in high-level semantics, which means that the inputs X and the targets Y probably have no dependency nor relation except for the high-level features $\tilde{X}$ they shared.

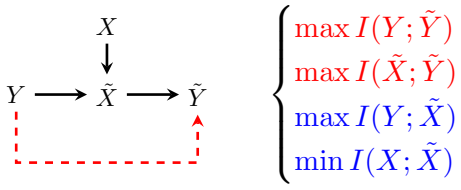

Figure 3.6: Information flow and optimization objective of GAN models for image-to-image translation

3.3 Information-theoretic Frameworks of Generative Models for Image Restoration

Now that we have analyzed the information-theoretic frameworks of different deep generative models, let's come back to discuss the relevant models for the tasks of image restoration.

3.3.1 Information Dynamic in the Models using the Conventional Idea of Image Restoration

The conventional understanding of image restoration tasks considers all information required for restoring the clean background images can be retrieved from their input images of observation ($I(Y;X) = H(Y)$), and therefore, the processes of image restorations models are regarded an information extraction process that extracts corresponding information from the source inputs.

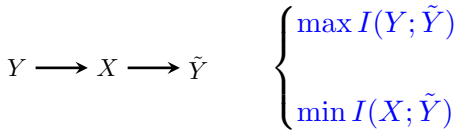

Figure 3.7: Information flow and optimization objective of models using the conventional idea of image restoration

Thus, models using the above idea for image restorations can be directly explained using the initial IB theory: for a visually-degraded image X to be fed into a restoration network, its desired output Y is the corresponding clean image of its original background scene, which, in reverse, determines the basic information of X. Suppose we consider the layers of the restoration network as a whole, hence we define the representation obtained from the latent layers as $\tilde{X}$ and the final outputs from the generator network $\tilde{Y}$ as the estimated restored image in approximation to Y. Their dependency relationship can form a Markov Chain: $Y \to X \to \tilde{X} \to \tilde{Y}$, where the optimization goal of the learning process is to maximize the mutual information between $\tilde{Y}$ and the ideal output Y while minimizing the mutual information between $\tilde{Y}$ and the input X:

$$\min[I(X;\tilde{Y}) - \beta I(Y;\tilde{Y})] \tag{3.10}$$

where β is a positive Lagrange multiplier that trades-off between the two terms.

According to the Data Processing Inequality (DPI) Cover Thomas and Thomas Joy (1991), we have:

$$I(Y;X) \geqslant I(Y;\tilde{X}) \geqslant I(Y;\tilde{Y}) \tag{3.11}$$

where the first equality is satisfied if and only if $\tilde{X}$ is a sufficient statistic of X, which requires the encoder network to be powerful enough to pass all the mutual information $I(Y;X)$ from X to $\tilde{X}$; and, similarly, the second equality is satisfied if and only if $\tilde{Y}$ is a sufficient statistic of $\tilde{X}$, requesting the decoder network to pass the entire information to $\tilde{Y}$. In this way the targeted mutual information $I(Y;\tilde{Y})$ can be maximized to reach $I(Y;X)$.

3.3.2 Information Dynamic in the Deep Generative Models for Image Restoration

In the real situations of the image restoration tasks, however, the observed images X may not contain all the information required for restoring the targeted clean background Y ($I(Y;X) < H(Y)$). Particularly for the tasks like image dehazing and image deraining, due to serious distortion like blurring and covering, some pixels of background in the observed image may not be recovered by using only the information from this single input. In fact, most data-driven generative models tend to more or less "imagine" the missing contents based on external knowledge learned from other inputs and predictions calculated by the models' parameters Bengio, Yao, et al. (2013); Goodfellow et al. (2014).

Moreover, in the modern networks structures of deep generative models, information may pass through the generator networks in separate paths, where the Markov Chain above may no longer stand. Besides the features learned and extracted by the network parameters $\tilde{X}$, some information in the inputs X may be directly passed through the generator network and retained in the generated outputs without calculation or learning process. More specifically in the image restoration tasks, they can be the background pixels or fine-grained details that are not degraded or distorted. To better reconstruct them, some generator networks provide structures like skip connections to assist in passing this intact information directly to the generating part of the network. Thus, this part of the information may not be involved in the optimization objectives or should be considered separately in the information-theoretic framework.

In this study, we consider the information required for generating the outputs $\tilde{Y}$ in the image restoration tasks originates from three sources (Fig. 3.8):

1. information provided by the extracted features $\tilde{X}$ from the feature extraction model or the encoder network: $I(\tilde{X};\tilde{Y})$;

2. information in source inputs X that pass directly through the skip connections or retained intactly for generating $\tilde{Y}$: $I(X \mid \tilde{X};\tilde{Y})$;

3. information involved by the parameters of the decoder network when calculating upsampling or interpolation: $H(\tilde{Y} \mid X,\tilde{X})$.

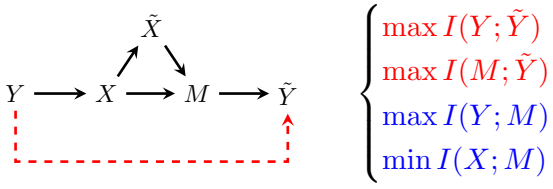

Figure 3.8: Information flow and optimization objective of GAN models used as generative models for image restoration

The Markov Chain $Y \to X \to \tilde{X} \to \tilde{Y}$ above no longer stands, but we can still analyze the information flow in the similar way: suppose the total information feed into the decoder network is $H(M)$, the learning goal become optimizing both M and $\tilde{Y}$ respectively:

$$\min[\beta_1 I(X;M) - \beta_2 I(Y;M) - \beta_3 I(M;\tilde{Y}) - \beta_4 I(Y;\tilde{Y})] \tag{3.12}$$

where maximizing the mutual information $I(Y;\tilde{Y})$ now becomes maximizing the mutual information of all three sources:

$$maxI(Y;\tilde{Y}) = max[I(Y;\tilde{X};\tilde{Y}) + I(Y;X \mid \tilde{X};\tilde{Y}) + I(Y;\tilde{Y} \mid X,\tilde{X})] \tag{3.13}$$

Ideally, in image restoration tasks, the three sources of information is to approach the optimal allocations as follow (See Appendix A for relevant proof and calculation):

$$\begin{cases} I(\tilde{X};\tilde{Y}) & \to -H(X \mid Y) \\ I(X \mid \tilde{X};\tilde{Y}) & \to H(X) \\ H(\tilde{Y} \mid X,\tilde{X}) & \to H(Y \mid X) \end{cases} \tag{3.14}$$

where the first case pursuits the encoder network to be powerful enough for extracting all information from X that is irrelevant to Y, which should be the patterns of image degradations like noise, haze, and rain; the second case requires intact information in X can be passed to $\tilde{Y}$, and the last case makes demands on the decoder to learn extra information so as to complete the missing information of Y in X as plausible as it can.

3.4 Review & Summary

To summarize, in this chapter, we extend the Information Bottleneck (IB) Principle to explain the Deep-learning-based Generative models (DGMs) and proposed a general information-theoretic framework for analyzing the generative methods for image restoration tasks. We conducted our analysis on the following aspects:

1. we analyzed three typical examples of general GAN models (original GAN, Conditional GAN, and GAN for image-to-image translation) by illustrating the information flow for

each of these models as well as their corresponding optimization objectives from the information-theoretical aspects respectively (Section 3.2);

2. we pointed out that: the conventional idea tends to consider the process of image restoration in the DGMs as a simple Markov Chain consisting of only a sinlge flow of information that tries to extract information about the targets from the source inputs (Section 3.3.1);

3. different from the conventional understanding above, we identified that in GAN models, the information involved in generating the restoration results actually comes from three sources: in addition to the information directly extracted by the network, some information in the inputs may pass intactly throughout the network without extraction process, and some external information that is not contained in the inputs may also be involved (Section 3.3.2);

4. inferred from the earlier analysis of the general GAN models, we obtain the information flow and corresponding optimization objectives of GAN-based models in image restoration tasks (Section 3.3.2);

5. based on the above framework, we calculated the optimal condition of information for each information component as well as the optimization boundaries for different parts of the generator networks, which reveals the learning behaviors inside the GAN models and can instruct the future design and analysis of DGMs for image restoration tasks (Section 3.3.2).

Chapter 4

Problem Formulation & Analysis

Although many conventional deep generative models (especially GANs) for image-to-image translation can be directly applied to the tasks of image restorations by considering them as end-to-end generation processes, they often fail in achieving satisfactory performance on quantitative results, compared with those traditional methods based on complicated handcrafted deconstructive models. Since GANs were not originally designed for the tasks of image restoration, the direct application of GANs in many existing image restoration methods tends to suffer from various issues and defects, leading to these performance gaps.

In this chapter, we will identify and elaborate on these issues when directly applying general GAN models to the image restoration tasks. According to the sources of information analyzed in the last chapter, we figure out and formulate three key problems, which are: over-invested abstraction, inherent loss of details as well as the vanishing gradients, and imbalance during the training process. Respectively, we try to explain these problems based on the information-theoretic framework above.

4.1 Over-invested Abstraction Process

4.1.1 Problem Definition & Formulation

Problem 1. *Patterns of the visual degradations in the image restoration tasks are locally distributed and relatively lower-level features, while generators designed for conventional GANs problems involve excessive abstraction processes for learning global and higher-level features, which may not be helpful for learning the features of image degradations when applied to the image restoration tasks, contributing to excessive network parameters, irrelevant information involved, and even drops in the restoration performances.*

Assumption 1. *Given the number of down-and-up-sampling layers in the conventional generator networks as $N_{convention}$, for a task of image restoration, there exist a $N_{saturated} < N_{convention}$, where continue increasing $N \leq N_{saturated}$ will not help with the restoration performance of the model or may even lead to performance drops.*

4.1.2 Intuition & Observation

Intuitively, we noticed that GANs were originally designed to learn abstract features for semantic-level generation tasks, whereas the patterns of visual degradations (like noise, haze, and rain)

to be learned in the image restoration tasks tend to be less abstract. In the traditional generation problems, GANs are to generate data that conform to the target distributions in high-level semantics (for example, drawing a variety of human faces), where the features to extract are often general and highly abstract (like the characteristics that can determine plausible faces). By contrast, in the image restoration tasks, the features of the image degradations tend to be much simpler, which can be regarded as relatively lower-level features according to Marr's definition Marr (1982). Moreover, unlike the conventional features that globally span large pixel areas, these patterns of image degradations tend to be sparse and locally distributed on the images. Therefore, the original design of the GANs models is likely to involve excessive abstraction processes when directly applied to the tasks of image restoration.

4.1.3 Analysis & Theoretical Explanation

More specifically, in the image-to-image translation models, this process of abstraction is often achieved by the down-and-up-sampling mechanism. The majority of generator networks are of encoder-decoder structure, where the down-sampling process in the encoder compresses information to $H(\tilde{X})$ with lower dimensionality, enforcing the network model to extract summarized and condensed key information, hence high-level semantic features can be learned.

More specifically, in the image-to-image translation models, this process of abstraction is often achieved by the down-and-up-sampling mechanism. The majority of generator networks are of encoder-decoder structure, where down-sampling in the encoders is mainly used to: 1. compress information to enforce learning high-level semantic features; 2. enlarge the receptive field to learn global features that span large pixel areas; 3. introduce inequality in information to prevent learning identical mapping or collapsing into trivial solutions; and, 4. reduce the computation and speeding up training and inference. Since the patterns of degradations in the image restoration tasks are relatively lower-level features, excessive information compression may not learn further features with higher-level semantics to help with the restoration performance. Moreover, as the pixels of the degradations are locally distributed in the input images, they may be easily detached from the background by using only a few layers of down-sampling. As for the last two functions, shallow layers of down-and-up-sampling can still achieve the same purposes. Thus, restoration models may not benefit from this kind of structure and there may exist excessive layers of down-and-up-sampling when applied to the tasks of image restoration.

We analyzed that these excessive down-and-up-sampling may not help with the image restoration tasks or may even contribute to performance drops. Suppose the total amount of information required to describe the features of the image degradataions is limited to H_{degrad}, where $H_{degrad} \leq H(X \mid Y)$. We consider this information is mainly extracted and passed by the encoder network through $I(X; \tilde{X})$, occupying a certain proportion in $H(\tilde{X})$. Given the total number of down-and-up-sampling layers in a generator network as N and the corresponding amount of information can be passed through each of these layers as $H(\tilde{X})_n^N$ $(n \in 1, 2, ..., N)$. For a UNet-like generator network (Encoder-decoder network with skip connections connecting corresponding down-and-up-sampling layers on both sides), the total amount of information can pass through the encoder network: $H(\tilde{X})^N = \sum_n^N H(\tilde{X})_n^N$, which increase along N. When

$H(\tilde{X})^{N'} \geq H_{degrad}$, continue increasing N may no longer help to extract the features of degradations for further improving the performance of models on the image restoration tasks, causing excessive network parameters, and may even involve extra information of noise $H(\tilde{X} \mid X)$. But for a generator of encoder-decoder without skip-connection, the total amount of information can pass is limited by the bottleneck layers: $H(\tilde{X})^N = \min\{H(\tilde{X})_n^N \mid n \in 1, 2, ..., N\}$, which decrease along N. When $H(\tilde{X})^{N'} \leq H_{degrad}$, continue increasing N will contribute to drops in the model's performances due to less enough information can be passed. Therefore, we deem that for both kinds of generator network, there exist a specific number of down-and-up-sampling $N_{saturated}$ where continuing to increase N may do no good to the overall performance of the model in the image restoration tasks.

4.2 Inherent Details Loss

4.2.1 Problem Definition & Formulation

Problem 2. *Existing generator models designed for traditional tasks tend to inherently discard low-level information in both extraction and generation processes, which, in the tasks of image restoration, is mainly related to the retaining of background information and fine-grained details in the source inputs, and may contribute to severe distortion and irrecoverable details loss in the restored results.*

Assumption 2. *Suppose the total amount of information received by the decoder network as M, where $H(M) = H(\tilde{X}) + I(X \mid \tilde{X}; M)$. We consider that in the conventional generators' network structure, the amount of low-level information that can actually be passed to the decoder is poorly sufficient to contain intact remaining information of X: $I(X \mid \tilde{X}; M) \ll H(X \mid \tilde{X})$. Besides, the decoder networks are neither capable enough to retain all the received information in their generated outputs: $I(M; \tilde{Y}) \ll H(M)$. In the tasks of image restoration, with low-level information like the background and relevant details occupying large proportions, the inherent discard of low-level information above in the models (reduce $I(Y; X \mid \tilde{X}; \tilde{Y})$) may contribute to more serious information loss (increase $I(Y; X \mid \tilde{Y})$).*

4.2.2 Intuition & Observation

Conventional generation tasks allow diverse variations in the details of the generated results, but this can be fatal for the tasks of image restoration. In the traditional generation problems, GANs and other generative models are only required to generate data that share consistency with the targets in high-level semantics but pay less attention to their details and other low-level information (like the eye color and hair texture in face generation). In fact, most of these methods even encourage these variations in details so as to guarantee the robustness of the learned representations. Nevertheless, this can be a completely different story for the tasks of image restoration: they put more emphasis on the preservation and retention of the details information and are trying to restore the exact background scenes from the corresponding input images. Therefore, the neglect and discard of the generated details in conventional GANs models can

be a crucial problem for their performances in image restoration tasks.

This problem of details loss can be apparently observed in their generated results (Fig. 6.3). For example, when applying conventional GANs models to the image deraining task, most rain-free images generated tend to be visibly distorted, suffering from different extend of loss in background details. What noticeable is: even in the pixel areas with no rain, the resulted images tend to display inaccurately from their corresponding inputs.

4.2.3 Analysis & Theoretical Explanation

We analyzed that: this issue of details loss is not only determined by the decoder network's capability to retain relevant information in the generated results $I(M;\tilde{Y})$ but also restricted by the amount of low-level information passed to the decoder $I(X \mid \tilde{X}; M)$. For the decoder, its information loss can be obvious: as a generative problem, extra information introduced by its network parameters can be inevitable (which is also necessary for approximating the absent information $H(Y \mid X)$): $H(\tilde{Y} \mid X, \tilde{X}) \neq 0$ and $H(Y) = H(X) = H(\tilde{X})$. Thus, only a certain proportion of the information that the decoder receives ($H(M)$) can be retained in the generated results ($H(\tilde{Y})$). Whereas more essentially, considerable low-level information has already been discarded before passing to the decoder. The total information passed to the decoder ($H(M)$) consists of two parts: high-level information abstracted and extracted by the encoder ($H(\tilde{X})$) and low-level information passed directly from X through structures like skip connections ($I(X \mid \tilde{X}; M)$). Since $H(\tilde{X})$ is excessive in containing all high-level information (discussed in *Problem 1*), ideally, we suppose $I(X \mid \tilde{X}; M)$ is maximized to contains all the information of $H(X \mid \tilde{X})$, so that intact information of X ($H(X) = I(X;\tilde{X}) + H(X \mid \tilde{X})$) can be achievable by the decoder. However, we noticed that the network structure of existing generator models does not support passing intact low-level information to the decoder without occupying $H(\tilde{X})$, even with skip connections: $I(X \mid \tilde{X}; M) \ll H(X \mid \tilde{X})$. Altogether, these two sources of information loss ($H(X \mid M)$ and $H(M \mid \tilde{Y}; X \mid \tilde{X})$) constitute total loss of low-level information in the generated results $\tilde{Y}$ ($H(X \mid \tilde{X}, \tilde{Y})$). Noticeably, in the image restoration tasks, the observed inputs X share large proportions of pixels about background and relevant details with the target outputs Y ($I(X;Y)$ is much larger than the other generative problems), which, we consider, are mainly low-level information. As a consequence, this inherent discard of low-level information in the generators tends to be more fatal in the restoration tasks, contributing to a more serious loss of details and distortion of the background scenes in the generated results (lower $I(Y; X \mid \tilde{X}; \tilde{Y})$ thus larger $I(Y; X \mid \tilde{Y})$).

4.3 Vanishing Gradient & Imbalance Training

4.3.1 Problem Definition & Formulation

Problem 3. *The source inputs in the image restoration tasks tend to contain too much information about the target outputs, where conventional measures of GAN loss may not provide smooth gradients for the continuous convergence of generative models and may result in imbalances*

between the generator network and the discriminator network during the training process.

Assumption 3. *In the tasks of image restoration, we consider* $I(Y;X) \gg H(Y \mid X)$*. Thus, conventional measures of loss optimizing* $\max I(Y;\tilde{Y})$ *tend to have large gradient on its component objective* $\max I(Y;X;\tilde{Y})$*, but fail to provide efficient gradients to* $\max I(Y;\tilde{Y} \mid X)$*, preventing the model from learning more generative knowledge for completing the missing information* $H(Y \mid X)$*. In GANs, this may also contribute to the imbalance between the generator and the discriminator where the generator tends to converge much earlier.*

4.3.2 Intuition & Observation

Compared with many other generation tasks, generated results in image restoration tend to show exceedingly high consistency with their targets, while in the traditional generation problems, this situation usually occurs only when the model is over-fitted. Thus, it can be intuitive to the concern that: with this high similarity, conventional measures of generation loss may no longer discriminate between them so well nor provide efficient gradients for the further optimization of the generated results.

4.3.3 Analysis & Theoretical Explanation

Reasons behind this high consistency may originate from the input information: in the tasks of image restoration, the inputs to the generator contain far more information about the ideal target Y than most of the other generation tasks. In traditional generation tasks, inputs X to the generator are often random noise Goodfellow et al. (2014); Radford et al. (2015) that is independent of the targeted outputs Y ($I(X;Y) = 0$), or short labels of condition Mirza and Osindero (2014) that does not involve much information ($I(X;Y) \approx 0$). Even for image-to-image translation Isola et al. (2017); J.-Y. Zhu, Park, et al. (2017), these generation tasks are conventionally cross-domain (like transferring from sketches to color images), in which the inputs (source domain) and the outputs (target domain) tend to differ considerably from each other. While in the image restoration task, by contrast, the input degraded images and the output restored images tend to be highly similar. In many cases, they even share the vast majority of the same pixels. Hence we can speculate that the distributions of the inputs and the outputs are likely to be close to each other and may even have $I(X;Y) > H(Y \mid X)$.

Existing measures that evaluate the loss in $\tilde{Y}$ for optimization is by directly calculating the pixel-wise similarities (L1-norm or L2-norm) or statistical divergence (KL-divergence or JS-divergence) between $\tilde{Y}$ and Y. This enforces and penalizes the entire generator model to optimize the mutual information between $\tilde{Y}$ and Y. According to the information flow of $\tilde{Y}$, it can be divided into two parts of optimization objectives:

$$\max I(Y;\tilde{Y}) = \max \underbrace{I(Y;X;\tilde{Y})}_{\mathcal{L}_1} + \max \underbrace{I(Y;\tilde{Y} \mid X)}_{\mathcal{L}_2} \tag{4.1}$$

For conventional generation tasks that $I(X;Y) \approx 0$, the first part of optimization objectives $\mathcal{L}_1 = I(Y;X;\tilde{Y})$ is often zero or negligible, and the optimization of the above measures tends

to be only maximizing the second term $\mathcal{L}_2 = I(Y;\tilde{Y} \mid X)$. Nevertheless, for the image-to-image translation in the image restoration tasks, the input images X contain far more information about target Y than most of the other generation tasks: $I(X;Y) > H(Y \mid X)$, which makes $\tilde{Y}$ easy to approximate Y by utilizing this information from X, where objective $\mathcal{L}_1$ converges much faster than $\mathcal{L}_2$ (expected gradient $\mathbb{E}\nabla\mathcal{L}_1 > \mathbb{E}\nabla\mathcal{L}_2$). As a consequence, conventional measures above may result in smaller values or even fail to provide gradients for further improving the generated results (gradient vanishing). For GAN models, these measures of performance may lead to an imbalance between the generator and the discriminator model since the generator may converge more easily by maximizing the information from X, while the discriminator needs to be trained with more data and epochs. Moreover, since the generated results $\tilde{Y}$ utilizing the information from X may have close distribution to the target outputs Y, it can be difficult for the network models to learn a decision boundary for distinguishing them, and improve the instability of GAN training.

Chapter 5

Proposed Solutions

We have analyzed that conventional GAN models may not be directly applicable to the tasks of image restoration due to over-invested abstraction, loss of details, as well as vanishing gradients, and imbalance in training. Whereas, these problems can be prevented or mitigated by modifying based on existing methods.

In response to the above issues, in this chapter, we analyze and discuss different potential approaches to cope with the problems above, such as modifying the network structure to optimize its capability in learning relevant features, enhancing network modules for details extraction, accumulation, and retention, as well as replacing the measures of the training objectives for more stable convergence. Ultimately, we come up with corresponding solutions and suggestions to improve GANs models' performances on the image restoration tasks.

5.1 Solutions for the Over-invested Abstraction Process

To prevent the over-invested abstraction process, a shallow network with fully-convolutional down-sampling and skip connections can be adopted as the backbone generator for the image restoration tasks. Here we consider a shallow U-Net Isola et al. (2017); Ronneberger et al. (2015) with $N_{saturated}$ layers of down-and-up-sampling is sufficient to extract most of the high-level information required.

5.1.1 Shallow Encoder Network

Since the abstraction process is often achieved by the down-and-up-sampling in the generator network, the most intuitive way to prevent the over-invested abstraction is to simply use a shallow encoder network by removing these excessive network layers of down-and-up-sampling. We have earlier analyzed in detail why removing excessive down-and-up-sampling layers can work: in the image restoration tasks, further information compression in these layers may not help the encoder to learn or extract higher-level features than the image degradations (like noise, haze, and rain), continue increasing the receptive field may neither be useful for extracting globally distributed information, while other functions of the down-and-up-sampling can still be achieved with shallow layers.

As we have analyzed earlier, for each image restoration task, there exists a minimum number of down-sampling layers $N_{saturated}$ that is enough for the neural network models to extract all features of the image degradations required for the task. Therefore, the practical measure is

to find the specific value of $N_{saturated}$ for the corresponding task and make its network layer number consistent with this value when constructing the network model.

5.1.2 Fully Convolutional Down-sampling

In addition to the number of network layers, the implementation of down-sampling in each of these layers may also affect the abstraction process. The conventional approach is to use pooling operation, which down-sampling by summarizing the features in patches of the feature maps with their average presences (average pooling) or the most activated presences (max pooling), using the idea of local translation invariance. However, these pooling-based downsamplings may not be suitable for tasks like the single image deraining: the exact pixel locations of the rain patterns within the patch will be obfuscated. By contrast, parameter-based down-sampling like fully-connected or fully convolutional layers may retain this fine-grained location information to the best extent. Comparatively, fully-connected layers involve more parameters than the fully convolutional layers and may require more data for training. Thus, we propose to use the fully convolutional layer as down-sampling.

5.1.3 Skip Connections

The skip connections between the encoder and the decoder in each down-and-up-sampling layer allow more information to pass throughout the network without being restricted by the bottleneck layer. It prevents information loss caused by over-compression along with the increase of layers. Moreover, these skip connections also sever the purpose of dissociating features in different level of abstraction, ensuring the down-and-up-sampling of the encoder network only focus on extracting high-level features, whereas lower-level features can be passed directly through these skip connections. Therefore, the excessive abstraction process will not cause loss to either information, ensuring that the overall performance of the model is not affected.

5.2 Solutions for the Inherent Details Loss

To reduce inherent details loss, we need to handle the discard of low-level information both before the decoder ($I(X \mid \tilde{X}; M) \ll H(X \mid \tilde{X})$) and inside the decoder networks ($I(M; \tilde{Y}) \ll H(M)$).

Noticably, we cannot directly increase the amount of low-level information that passed to the decoder $I(X \mid \tilde{X}; M)$ by simply modifying the skip connections without affecting $H(\tilde{X})$. As an alternative, we proposed to increase the total amount of information in the inputs to achieve this goal. More specifically, we propose a network module named Dilated Dense Block (DDB) (Fig. 5.2) to enhance the extraction and accumulation of low-level features in the inputs, where the number of layers in this DDB module can be used to reflect the total amount of low-level information passed to the decoder network. Utilizing the benefit of skip connections that all information from the previous layers can be retained in its outputs, we refer to the DenseNet Huang et al. (2017) structure for concatenating feature maps and accumulating information.

Moreover, to enhance the extraction of features in the same level and get rid of the effect caused by the receptive field, we also refer to the Dilated Convolution F. Yu and Koltun (2015) in the design of the DDB module. We applied the DDB modules to the inputs images before passing them to the backbone generator network.

To alleviate details loss that happens inside the decoder, we also proposed to enhance the decoder network to better retain information in their generated outputs. Referring to the generative methods of super-resolution for generating higher quality images, we modified the up-sampling process of existing decoder networks by adopting the sub-pixel convolution Shi et al. (2016) (SCU) as the enhancement.

5.2.1 Broadened & Global Skip Connections

We have pointed out that the discards of low-level information before the decoders is one of the reasons contributing to the details loss, which is inherent in the network structure of the generator models: $I(X \mid \tilde{X}; M) \ll H(X \mid \tilde{X})$. For the encoder-decoder generators without skip connections, this is obvious: only the abstracted high-level information in $H(\tilde{X})$ is passed to the decoder while all low-level information is discarded before passing to the decoder ($H(X \mid \tilde{X}) = 0$). Whereas, for UNet-like generator networks, skip connections are used to allow direct transmission of a certain amount of information in each abstraction level to the decoder, but there is no skip connection to directly connect the pure information in the original source inputs without abstraction or calculation to the decoder networks.

An intuitive idea is to enhance the skip connections in these generator networks in a bid to increase the amount of low-level information that can be passed to the decoder networks. More specifically, two rough measures can be applied: (i) broaden the existing skip connections in UNet generator Isola et al. (2017); Ronneberger et al. (2015) to allow more low-level information to be passed, or (ii) introduce a global skip connection that connects the inputs channels directly to the top layer of the decoder network so that intact information of inputs is accessible by the decoder. Nevertheless, we noticed that the former idea will unavoidably increase $H(\tilde{X})$, and the latter one will put pressure on the decoder network. We conducted relevant experiments and the results show that these naive measures do not bring significant improvement to the baseline models.

5.2.2 Dilated Dense Block

Rather than directly increasing the information in these skip connections $H(X \mid \tilde{X})$, we proposed to increase the total amount of information in the inputs $H(X)$ as an alternative solution. For the low-level information we intend to enhance, there is:

$$H(X \mid \tilde{X}) = H(X) - I(X; \tilde{X}) \tag{5.1}$$

Since $I(X; \tilde{X})$ is supposed to be the features of the image degradations, we consider it to be constant. Therefore, we can simply increase the amount of information of inputs $H(X)$ before

sending them to the generator networks to indirectly increase $H(X \mid \tilde{X})$ without modifying the skip connections or network structure of the generator network.

More specifically, we proposed to introduce a network module that can enhance the extraction and accumulation of information before sending them to the generator network. We refer to the network structure of Densely Connected Network (DenseNet) Huang et al. (2017): by using concatenative skip connections, feature maps in the previous layer can be reused in the deeper layers of the network. Thus, source information from the inputs can be fully retained and repeatedly emphasized for further extraction. For a given input $\mathbf{x}_0$, the output of a Dense Block can be represented as a recursive concatenation of L layers:

$$\mathbf{x}_l = concat([\mathbf{x}_{l-1}, F_l(\mathbf{x}_{l-1})]) \tag{5.2}$$

where $F_l(\cdot)$ denotes the operations in dense layer l.

Notably, by considering the outputs from all L layers as a whole, $\Psi(\mathbf{x}_0)$, where $\Psi(\mathbf{x}_0) = concat([F_1(\mathbf{x}_0), F_2(\mathbf{x}_1), \ldots, F_l(\mathbf{x}_{l-1})])$ represents a concatenation of extracted feature maps from each layer, the entire outputs of this kind of structure can be regarded as a concatenation of the input $\mathbf{x}_0$ and these extracted features $F_l(\mathbf{x}_{l-1})$ from each layer: $\mathbf{x}_l = concat([\mathbf{x}_0, \Psi(\mathbf{x}_0))])$. It indicates that the original input $\mathbf{x}_0$ is preserved in its entirety through a direct connection from the beginning to the end, where the later processes can still have intact information of the original source input.

The Residual Network (ResNet) He, Zhang, Ren, and Sun (2016) also has a similar network structure by using skip connections to pass information to deeper layers. However, it achieves in an additive manner, which applies in-place addition of the learned residual features with the layer's input. Therefore the output feature maps may hardly contain intact input information for later processing.

Zhang et al. H. Zhang et al. (2019) also adopted the DenseNet structure in their GAN-based deraining model. However, instead of placing the DenseNet module before the down-sampling processes to emphasize the input information, it applies dense blocks after the pooling layers of the network, where details information might have already been lost in the foregoing down-sampling process. Figure 5.1 illustrates the difference between the previous model and our proposed network.

Furthermore, we consider some fine-grind details within the patches of the convolutional filters may be obfuscated and hardly recovered if all filters are of the same size. To help with the extraction of these features and to eliminate the interference caused by the difference in receptive fields, we adopt the idea of multi-scaling, so as to aggregate contextual information from different receptive fields. More specifically, we refer to the Dilated Convolution F. Yu and Koltun (2015) to obtain a larger receptive field without increasing the number of layers or involving extra parameters and achieve the above idea by using a multi-path structure, concatenating convolutions with different dilation rates.

Our proposed dilated dense block (DDB) is indicated as follow:

$$\mathbf{x}_l = concat([\mathbf{x}_{l-1}, F_l(\mathbf{x}_{l-1}), G_l(\mathbf{x}_{l-1}), H_l(\mathbf{x}_{l-1})]) \tag{5.3}$$

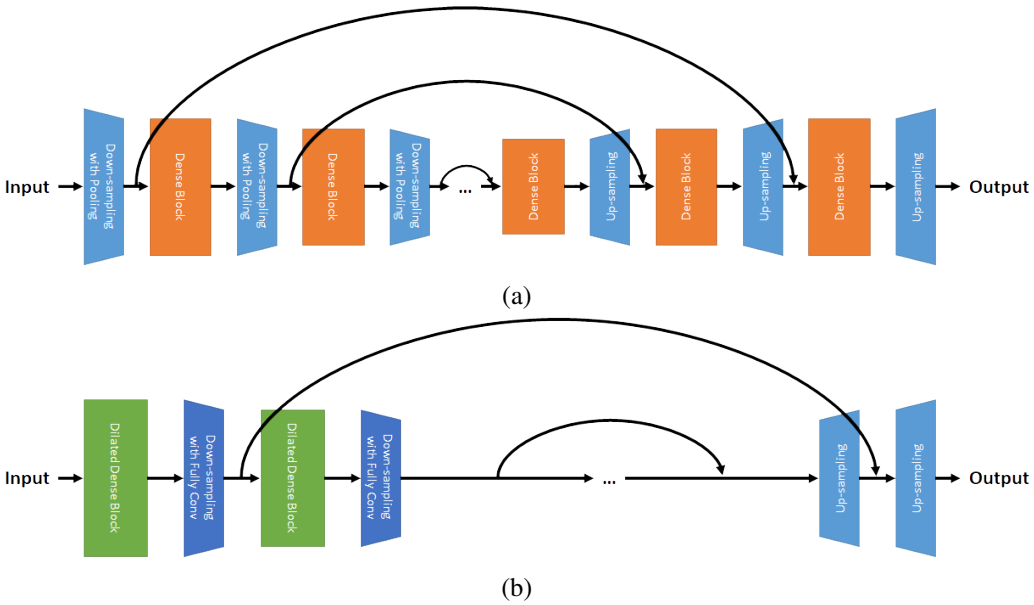

Figure 5.1: Comparison between the generator network of ID-CGAN with DenseNet structure and our proposed detail-enhancing generator using DDB module(s).

with $F_l(\cdot)$, $G_l(\cdot)$ and $H_l(\cdot)$ represents the composite functions involving convolution with dilation rate 1, 3, and 5 respectively.

Due to the reuse-ability of features, each dense layer only needs to focus on extracting a small number of features, and the overall feature extract-ability of the dilated dense block can be determined only by the number of dense layers inside. Theoretically, the number of feature maps in the output is related to both the growth rate and the number of layers. But in this case, a larger growth rate is equivalent to adding extra layers, because the inputs to all the dense layers include complete source data and thus the features extracted from each layer are independent. Therefore, we simply assign a relatively small value to the growth rate and determine the complexity of the feature to be extracted by adjusting only the number of layers, so as to control the feature extract-ability.

5.2.3 Sub-pixel Convolutional Upsampling

The loss of details also exists in the up-sampling process of the decoder network. The earliest up-sampling methods based on un-pooling (missing pixels are abandoned), or interpolation operations (missing pixels are filled based on their neighbors) involve irreversible information loss. Better solutions try to fill the missing pixels with spatially-adjacent textures, or contextual information. For example, in the UNet of the pixel2pixel model, up-sampling is achieved using deconvolution (transposed convolution), which is useful for involving some more general information when filling the missing pixels. However, all these up-sampling methods do not retain the input details and try to fill the missing pixels with calculated results, which is likely to introduce noises or information that is inconsistent with the source inputs, or contributes to the Checkerboard Artifacts Aitken et al. (2017) in the generated results.

Sub-pixel convolution Shi et al. (2016) is a better solution for up-sampling, which is com-

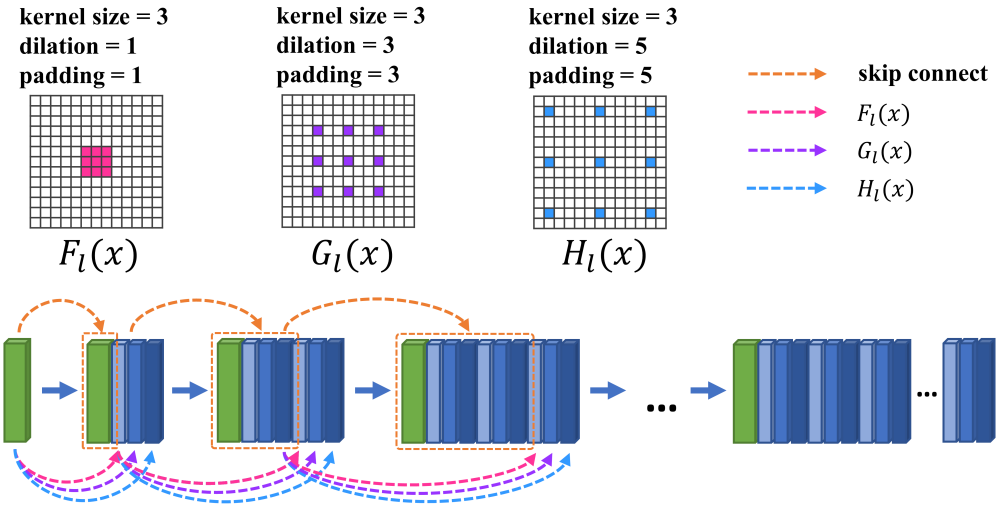

Figure 5.2: Proposed network module of Dilated Dense Block (DDB)

monly used in applications like super resolution for generating higher quality images. A sub-pixel convolution module often consist of a convolution layer and a pixel-shuffle operation, in which an input of $H \times W \times Cr^2$ tensor will be rearrange to form a $rH \times rW \times C$ tensor using phase shift (r denotes the upscale factor):

$$\mathcal{PS}(T)_{h,w,c} = T_{\lfloor h/r \rfloor, \lfloor w/r \rfloor, c \cdot r \cdot mod(w,r)+c \cdot mod(h,r)+c} \tag{5.4}$$

where h, w and c corresponds to the height, weight and number of channels in the resulted image.

To retain details in the generated images to the largest extent and prevent the Checkerboard Artifact, we proposed to use the sub-pixel convolution up-sampling (SCU) at the top layer of the decoder network.

5.3 Solutions for the Vanishing Gradient & Imbalance Training

On account of the vanishing gradient and imbalance during the training process, we proposed to use the LSGAN loss Mao et al. (2017) to replace the traditional GAN loss that is based on JS-divergence Goodfellow et al. (2014). It can reduce the problem of vanishing gradient when the distributions between the targets the generated results are fairly close to each other, and allow further convergence of GAN models. In addition, another helpful suggestion to cope with the imbalance of training between the GANs' networks is to pre-train the discriminator using extra data.

5.3.1 Least Square GAN Loss

As we analyzed earlier: generators of GAN models applied to the image restoration tasks can easily converge utilizing the relevant information provided by the inputs X to approximate the

targets Y, which may lead to the fairly close distributions between Y and the generated outputs $\tilde{Y}$ at the early stage of training. In this case, the discriminators with less sufficient training may hardly distinguish between the generated results and the real data samples, contributing to the vanishing gradient and imbalance between the generators and the discriminators.

To allow further convergence of the models and prevent early stopping of training, we propose to apply the Least Squares GAN (LSGAN) loss based on Pearson χ^2 divergence in replace of the traditional Binary Cross-Entropy (BCE) loss based on JS-divergence, which ensures that the loss function can still provide valid gradients when the distributions between $\tilde{Y}$ and Y are fairly close to each other (Fig. 5.3). In traditional GAN models, the training objectives are measured with the JS-divergence between the two distribution, whose dynamic lies on a sigmoid-like function. In the case of image restoration, where the two distributions (source and target) are close and may have a large proportion of overlapping, the fitted loss function curve can be flat and less directive. Moreover, the noise introduced by the generator can even weaken the gradient of these target functions. LSGAN uses a linear function to represent the cost function of training, which provides constant gradients as long as the two distributions measured are not fully over-lapped. Since the noise introduced by the generator models is always less significant than the actual difference between the targets and the generated results, the linear function can always point in the right direction.

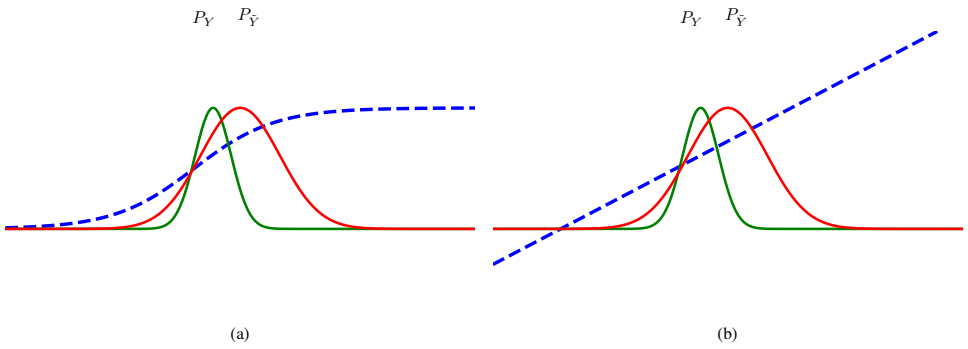

Figure 5.3: A top example comparing the BCE loss used in original GAN and LSGAN loss. In case the two distributions are close and one of them has bi-directional biases, the gradient of the loss function in the original GAN tends to be flat and can be less instructive. But LSGAN loss can always have non-zero gradients (least-squares tend to have linear derivatives), as long as the two distributions are not fully overlapped. Moreover, because the biases caused by noises are often smaller than the actual biases between the source and the targets, the linear function of the average gradient can generally slant in the right direction, ensuring the gradients always exist along the whole training process.

5.3.2 Discriminator Pre-training Using External Data

Since in the GAN models, the generators tend to converge much earlier than the discriminators using the same amount of training data, to enhance the discrimination ability of the discriminator, extra training can be applied as the pre-training of the discriminators.

Chapter 6

Experiments & Results

In the former parts of this thesis (Chapter 4), we indicate the three key problems (over-invested abstraction process, inherent details loss, as well as vanishing gradient, and imbalance in training) that exist when directly applying deep generative models to the tasks of image restoration and formulate respectively them as hypotheses. Correspondingly, we discussed and proposed solutions and methods for each of these problems in Chapter 5, trying to improve the performances of deep generative models on image restoration tasks.

In this chapter, we provide empirical evidence in proof of the above problems and present results of the corresponding contrast experiments to validate the proposed solutions and methods. Here we focus on the image deraining and dehazing tasks in image restoration, conducted relevant experiments, and reveal corresponding conclusions mainly based on quantitatively measuring their performances on deraining / dehazing datasets.

6.1 Datasets

In this study, we conducted all training experiments and evaluated relevant models mainly on six benchmarking datasets of image restoration as follows:

- one image denoising dataset - *SIDD-sRGB* Abdelhamed, Lin, and Brown (2018)
- one image dehazing dataset - *RESIDE-ITS* B. Li et al. (2019)
- two image deraining datasets - *Rain800* H. Zhang et al. (2019) & *Rain12000* H. Zhang and Patel (2018)
- two datasets with rain and haze appear simultaneously - *RainCityScapes* Hu et al. (2019) & *OutdoorRain* R. Li et al. (2019)

To reduce the computational cost for training, we only use the smallest subset of *SIDD-sRGB* (i.e. *SIDD-Small-sRGB*) for training our models, but we evaluate our models on the entire benchmark of the *SIDD-sRGB* dataset (i.e. *SIDD-Validation-sRGB*). Similarly, on account of the huge amount of data in the *RESIDE-ITS* dataset, we randomly split with ratio 8:2 (11192:2798) as the training and testing set used in our experiments (we denoted as *RESIDE-ITS-8-2*). Moreover, Since the testing set of the *OutdoorRain* dataset is not yet public available, we randomly split its training set with ratio 8:2 (7200:1800) as our *OutdoorRain-8-2* datasets in this paper.

Noticeably, apart from the observed image inputs and their corresponding ground truths, three datasets above: *RESIDE-ITS*, *RainCityScapes* and *OutdoorRain* provide additional training data to provide extra supervision for their proposed methods. The *RESIDE-ITS* dataset provides the layers of haze for each input, the *RainCityScapes* dataset contains maps of scene depth for each training data, and the *OutdoorRain* provides the ground-truths of rain streak layers, atmosphere light layers, as well as transmittance layers as supervision. All this information is useless for generative models, and we only use the hazy(rainy) inputs and their ground truths for training and evaluation.

Detailed statistics for the datasets are summarized in Table 6.1.

Dataset	Image Restoration Tasks	# Training Data	# Testing Data	Composition Method
SIDD-sRGB	Image Denoising	160 (SIDD-Small)	1,280 (SIDD-Validation)	Linear Additive Composition
RESIDE-ITS-8-2	Image Dehazing	11,192	2,798	Atmospheric Scattering Model
Rain800	Image Deraining	700	100	Linear Additive Composition
Rain12000		12,000	1,200	Density-aware Additive Composition
RainCityScapes	Image Dehazing + Deraining	9432	1188	Depth-aware Composition
OutdoorRain-8-2		7200	1800	Heavy Rain Model + Depth-aware Composition

Table 6.1: Information of the datasets used in our experiments

6.2 Evaluation Metrics

We adopted the peak signal to noise ratio (PSNR) Horé and Ziou (2010); W. Yang, Tan, Wang, Fang, and Liu (2020b) and structural similarity index (SSIM) Brooks, Zhao, and Pappas (2008); Horé and Ziou (2010) as the quantitative methods to evaluate the performances of models on all image deraining, image dehazing and image reconstruction tasks involved in this thesis.

Definition 6.2.1 (PSNR)**.** Suppose X and Y are a reference image and a test image, both of size $M \times N$, the PSNR between X and Y is calculated by:

$$PSNR(X,Y) = 10\log_{10}(\frac{255^2}{MSE(X,Y)}) \tag{6.1}$$

where:

$$MSE(X,Y) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(X_{i,j} - Y_{i,j})^2 \tag{6.2}$$

Definition 6.2.2 (SSIM)**.** The SSIM between images X and Y is calculated by:

$$SSIM(X,Y) = l(X,Y)c(X,Y)s(X,Y) \tag{6.3}$$

and

$$\begin{cases} l(X,Y) = \frac{2\mu_X\mu_Y + C_1}{\mu_X^2 + \mu_Y^2 + C_1} \\ c(X,Y) = \frac{2\sigma_X\sigma_Y + C_2}{\sigma_X^2 + \sigma_Y^2 + C_2} \\ s(X,Y) = \frac{\sigma_{XY} + C_3}{\sigma_X\sigma_Y + C_3} \end{cases} \tag{6.4}$$

where C_1, C_2, and C_3 are positive constants used to avoid a null denominator; μ_X, μY, σ_X, σ_Y, and σ_{XY} are mean of X, mean of Y, standard deviation of X, standard deviation of Y and the convariance between X and Y respectively.

For both PSNR and SSIM, larger values indicate better performances of models.

6.3 Empirical Evidence for the Over-invested Abstraction Process

To prove the existence of over-invested abstraction processes, we can investigate the performances of GANs models with different levels of abstraction learning abilities. More specifically, we set up groups of GAN models with different numbers of down-and-up sampling layers in their generator networks, then we trained and compared their performances on the six different image restoration datasets above.

Here we adopted two common types of generator networks used for image-to-image translation as the backbones of GAN models: a convolutional encoder-decoder without skip connection Noh, Hong, and Han (2015) (denoted as *En/Decoder*), and a U-Net Ronneberger et al. (2015) (*UNet*). For the *En/Decoder* generator, 2×2 max-poolings are used as the down-sampling method, and nearest-neighbor interpolations with scaling factor of 2 are used for the upsampling processes. Convolutional layers with filter size of 3×3 and padding of 1 are also used for learning features in both the encoder and decoder parts of the network but without down or up sampling. *UNet* Ronneberger et al. (2015) is first introduced as a generator network in the pixel2pixel model Isola et al. (2017) with 8 layers by default. It uses fully convolutional layers for both down-sampling and up-sampling with skip connections concatenating outputs at each level. For both generator networks, the scaling factors are set to 2, and to be consistent, we compared the deraining performances of both models with 1 to 8 layers of down-and-up-sampling respectively.

The experiment results (Fig. 6.1 shows the corresponding results of deraining performances on the Rain800 dataset) indicate that: when the number of down-and-up-sampling layers N reaches a certain value, continuing increasing N does not improve the performance of the model (*UNet*) on image restoration tasks, or may even cause a performance drop (*En/Decoder*). It means that the extra abstraction process does no good to the tasks of image restoration. For the *En/Decoder* generators, the deraining performances of models sharply drop after a short climbing (at around 2 layers in Fig. 6.1) along with the increase of down-and-up-sampling layers. It tends to because the amount of information that passes throughout the networks is limited by their bottleneck layer, where too much information compression may not be able to pass enough information for restoring the targeted clean images. These results also reflect that considerable amounts of low-level information without abstraction process plays an essential role in the image restoration tasks. Whereas, for the *UNet* generators, the performances of the models do not suffer this kind of problem: when increasing the number of down-and-up-sampling layers, higher-level features can be extracted, while low-level information from the previous layers can still pass through the skip-connections in the networks. Noticeably, however, the performances of models no longer improve after the relevant number of layers reaches certain

values (for instance, the 4th layer in Fig. 6.1, where 4-layer U-Net can already achieve the same deraining performance as the 8-layer U-Net used in the pixel2pixel models). It indicates that the level of abstraction reaches its saturation here, where no higher-level feature can be extracted to help with the tasks, even if we continue deepening the U-Net generator networks. Excessive abstraction processes exist and may not be helpful for the tasks.

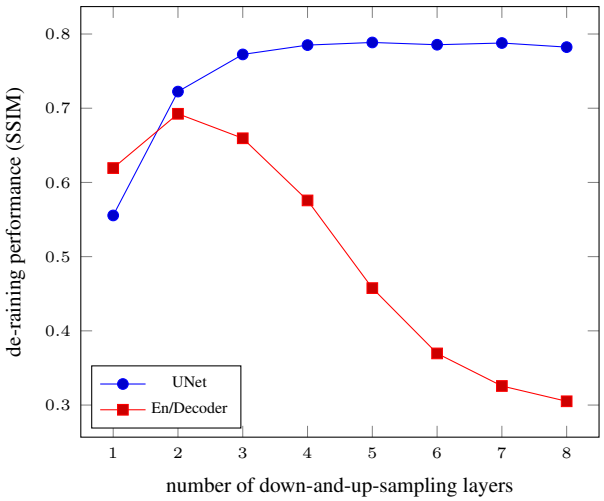

Figure 6.1: Deraining performances of GAN models using two different types of generator networks with different numbers of down-and-up-sampling layers

By comparing the relevant experimental results on different image restoration tasks and datasets (Fig. 6.2 compares the corresponding trends of image restoration performances of GAN models using U-Net generators on three datasets: SIDD, SIDD-sRGB, and Rain800, representing denoising, dehazing, and deraining tasks of image restoration respectively), the saturated numbers of the down-and-up-sampling layers tend to be varied on different datasets. It means that different levels of abstraction, as well as different amounts of information, may be required for different image restoration tasks. It can also be intuitive that the corresponding saturated numbers tend to be relatively low on the image denoising datasets, while models for image deraining and image dehazing datasets may require more down-and-up-sampling layers to reach their optimal performances since the features in image denoising tasks are likely to have much lower levels of abstraction than image deraining and dehazing tasks. Generally, for the 8-layer-UNet used in pixel2pixel Isola et al. (2017), we found that 5 layers of down-and-up sampling tend to be saturated on all six image restoration datasets we have tested.

6.4 Empirical Evidence for the Inherent Details Loss

The issues of inherent details loss in the deep generative models can be apparently observed in their generated results. Here we demonstrate the image deraining results generated by the conventional GAN model (pixel2pixel Isola et al. (2017)) on the real scene data and the relevant performances by the corresponding model after enhancement of its details extraction-ability.

Although the existence of details loss can be evident in the corresponding generated results,

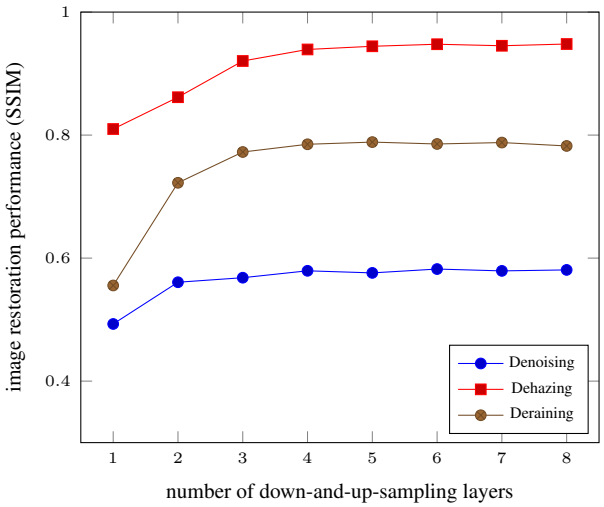

Figure 6.2: Performances of GAN models on different image restoration tasks using U-Net generator networks with different numbers of down-and-up-sampling layers

we also proved it quantitatively by training these generator models directly as auto-encoders (learn to perform reconstruction of the input images) (See Fig. 6.4). We input the clean background images (ground truth) from the Rain800 dataset to these generator networks and train them to output images that are as similar to their inputs as possible by introducing MSE loss between the inputs and outputs. The higher similarity between the inputs and the outputs indicates less information is lost in the generator network. We compare corresponding reconstruction performances of different generator networks. The results show that none of the generator networks can completely reconstruct the input images meaning that all of them more or less suffer from the problem of details loss. Noticeably, the model with details enhancement (methods proposed above) achieves the best restoration performance, with an average PSNR of the restored images reaching 50.0 and an average SSIM reaching 0.9990.

To prove that this issue of details loss originates from the discard of low-level information before the decoder, where improving the extraction and accumulation of information in the inputs can help, we investigate the deraining performances of models equipped with the DDB modules in different numbers of layers (Fig. 6.5). Since in the DDB modules (or DenseNet module Huang et al. (2017)), the number of dense layers can represent the accumulation of information passed to their later networks, by adjusting this number of layers in the network module before the generator network, we can control the corresponding amount of information about its inputs to be passed to the generator afterward. Here we study their corresponding performances on the Rain800 dataset, using a 5-layer-UNet as the backbone generator and adopting the DDB as the network module to enhance the inputs' information before being sent to the generator network.

Corresponding results indicate that: the overall performances of models on image deraining do improve along with the increase of layers in the DDB modules, meaning that increasing the amount of information before the generator networks does help the image deraining task. Since the actual amount of information inside the generator networks is restricted by their network

(a) input (b) pix2pix (c) pix2pix + DDB

Figure 6.3: Deraining results of pixel2pixel Isola et al. (2017) on real data with and without details enhancement

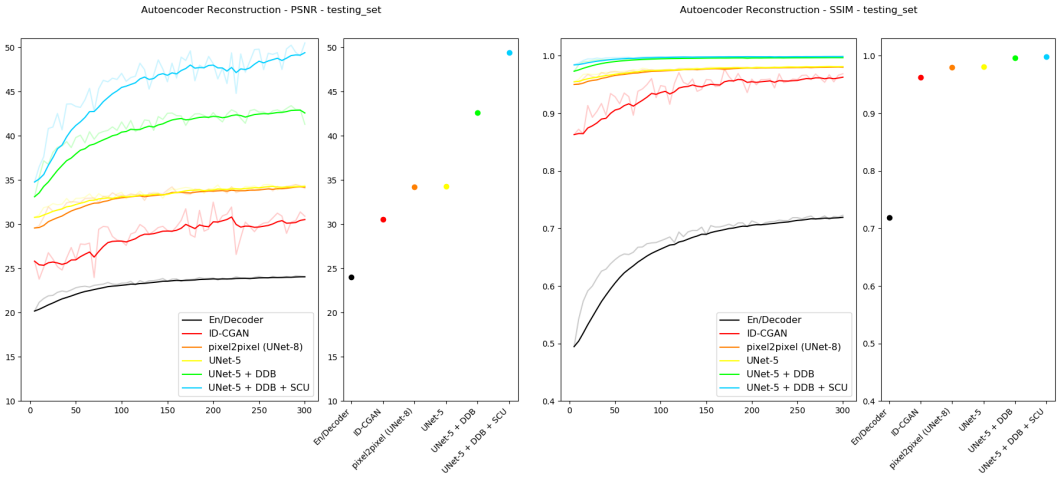

Figure 6.4: Reconstruction performance of different generator models.

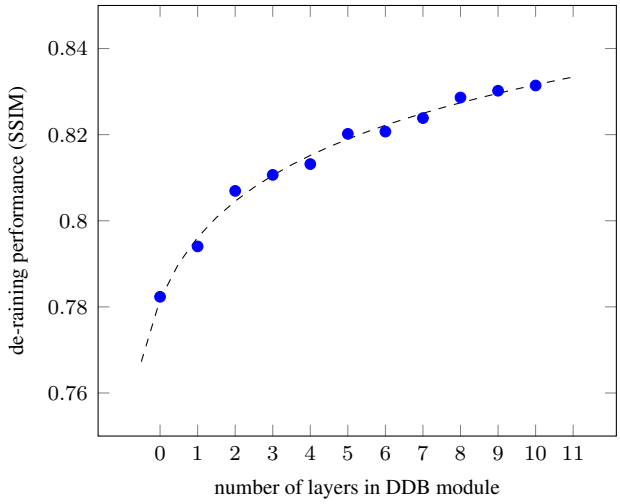

Figure 6.5: Deraining performances of GAN models equipped with DDB modules in different numbers of layers.

structures, increasing the information of their inputs does not affect the amount of information received by the generation process. Thus, it may only play the role of emphasizing certain information for better extraction or distinguishing in the later network. Similar results are also observed when using DenseNet module Huang et al. (2017), meaning that information accumulation is the functions that actully make difference.

To prove that the details loss also happens inside the decoder network, and enhancing the decoder network can help, we introduced the SCU module to the decoder network and compare its performances on both image deraining and image reconstruction with models without SCU (similar experiment settings as above). Results on decoders with SCU module indicate apparent advances on both deraining and reconstruction performances, meaning that enhancing the decoder network does help to alleviate the problem of details loss.

To demonstrate that the details loss also happens inside the decoder networks, we compared the relevant performances of models with and without SCU modules applied to enhance the decoder networks. Results on decoders with SCU module indicate apparent advances on both deraining (Table 6.2) and reconstruction performances (6.4), meaning that enhancing the decoder network does help to alleviate the problem of details loss.

6.5 Empirical Evidence for the Vanishing Gradient & Imbalance Training

The issue of vanishing gradient and imbalance training can be observed during the relevant training process: we trained the pixel2pixel Isola et al. (2017) model on four datasets, and in all trials, the discriminators converged much earlier than the generators, making the training hard to continue. All these trainings ended up with relatively large values of the generator losses, while the discriminator losses all tend to converge to zero. This is commonly regarded as a failure of training in GANs, where the discriminator fails to provide gradients for the generator to continue training.

We also observed a two-stage convergence (Fig. 6.6) in most of our experiments, where the loss functions during the training tend to converge fast in the first stage but suddenly slow down in the second stage. This seems to coincide with our earlier analysis, where objective $\mathcal{L}_1$ in Eqn. 4.1 converges much faster than $\mathcal{L}_2$, contributing to the two-stage convergence.

6.6 General Experiments

Generally, we evaluate the overall performances of models in image deraining and image dehazing by applying the proposed solutions above and compare with their corresponding baselines. More specifically, we applied a shallow 5-layer-U-Net the backbone generator, with a 15-layer-DDB module added before the generator network to enhance extraction and retaining of details information, SCU module added to the top layer of the decoder network to enhanced retaining of details, and adopt the LSGAN loss for training and optimization. We compared the relevant performance on four datasets. Relevant results are demonstrated in Table 6.2. Results indicates

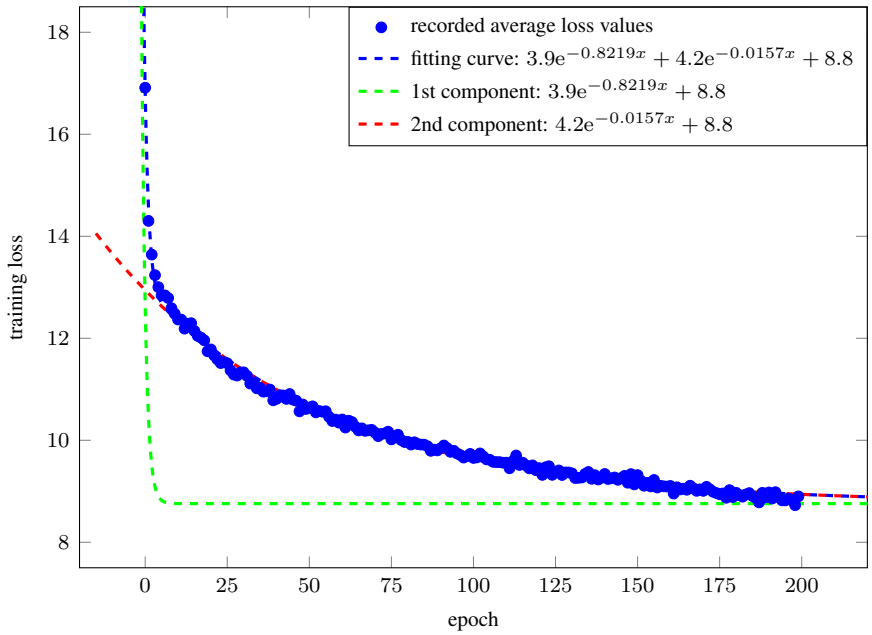

Figure 6.6: Variation of the GAN loss function along the training process

that the proposed solutions achieve apparent improvements on the backbones and outperform relevant baselines without the use of extra supervision from the dataset and with much lighter-weight models.

	Rain800		**Rain1200**		**RainCityScapes**		**OutdoorRain-8-2**	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DID-MDN H. Zhang and Patel (2018)	-	-	27.95	0.9087	28.43	0.9349	-	-
DAF-Net Hu et al. (2019) *	-	-	-	-	30.06	0.953	-	-
HeavyRainRestorer R. Li et al. (2019) *	-	-	-	-	-	-	22.12	0.7821
ID-CGAN H. Zhang et al. (2019)	23.98	0.8029	25.99	0.8022	-	-	-	-
pix2pix Isola et al. (2017) (UNet-8)	23.47	0.7645	26.34	0.7759	28.61	0.9345	25.45	0.7934
pix2pix Isola et al. (2017) (UNet-8) + LSGAN Mao et al. (2017)	24.12	0.8053	29.53	0.8719	30.92	0.9547	-	-
UNet-5 + LSGAN Mao et al. (2017)	24.09	0.8045	29.16	0.8745	30.88	0.9653	26.74	0.8564
UNet-5 + LSGAN Mao et al. (2017) + DDB	25.08	0.8116	30.77	0.9013	-	-	-	-
UNet-5 + LSGAN Mao et al. (2017) + DDB + SCU	25.29	0.8135	31.52	0.9053	31.29	0.9645	27.5	0.8799

$*$ indicates model that require extra information (ground-truth of depth map, ground-truth of rain streak layers and transmission map etc.) from the datasets as supervision.

Table 6.2: Evaluation results on deraining & dehazing datasets

6.7 Ablation Experiments & Supplemental Results

6.7.1 Positions of Adding the Detail Enhancing Module

Originally, we intend to apply the detail enhancing network module before the generator network to help extraction and accumulation of low-level information. We also investigate the

models' performances by applying the network modules on different positions of the generator networks (Fig. 6.7). Here we use the 8-layer-UNet in pixel2pixel Isola et al. (2017) as the backbone generator and try to insert 15-layer-DDB modules before each of its encoder layers ("1st" denotes adding a DDB module before the generator, while "1st - 8th" means that 8 DDB modules are added before all 8 encoder layers of the UNet generator). Similarly, we train the pixel2pixel model on the Rain800 datasets.

We observe that adding the DDB module to the "1st" position brings the greatest improvement while adding which to deeper layers does not may much different to the deraining performance of the model. This also reveals the inherent discards of low-level information in the network structure before the decoder network. Noticeably, adding extra DDB at the "2nd" position also make minor improvement on the models. It may indicate that some relatively higher-level information is also enhanced by the DDB module.

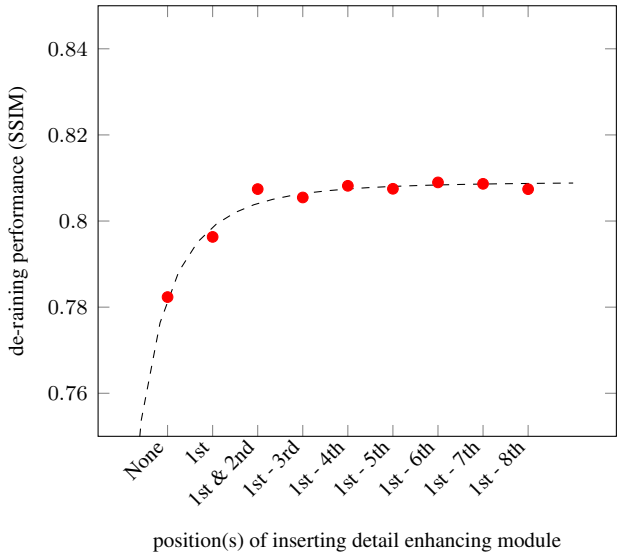

Figure 6.7: Deraining performances of pixel2pixel models Isola et al. (2017) with detail-extraction enhancing modules inserted to different positions of their backbone generator networks

6.7.2 DDB Modules Compared with other different Network Modules

Relevant studies have also proposed considerable network modules to enhance the deraining performances of their models. Here, we also compare our DDB with these modules (Fig. 6.8), including Residual deraining module (Residual) Fu et al. (2017), Contextualized Dilated Block (ContextDilated) W. Yang et al. (2017), SCAN module (SCAN) X. Li, Wu, Lin, Liu, and Zha (2018), Recursive deraining module (Recursive) Fu et al. (2019), Attentive Recurrent module (Attention) Qian et al. (2018), and ordinary DenseNet module (Dense), as contrast to the proposed DDB module. We use a 5-layer-UNet as backbone generator with SCU module and compare both their image deraining and image reconstruction performances. Results indicate that the proposed DDB module makes the greatest improvement on the baseline model than other network modules.

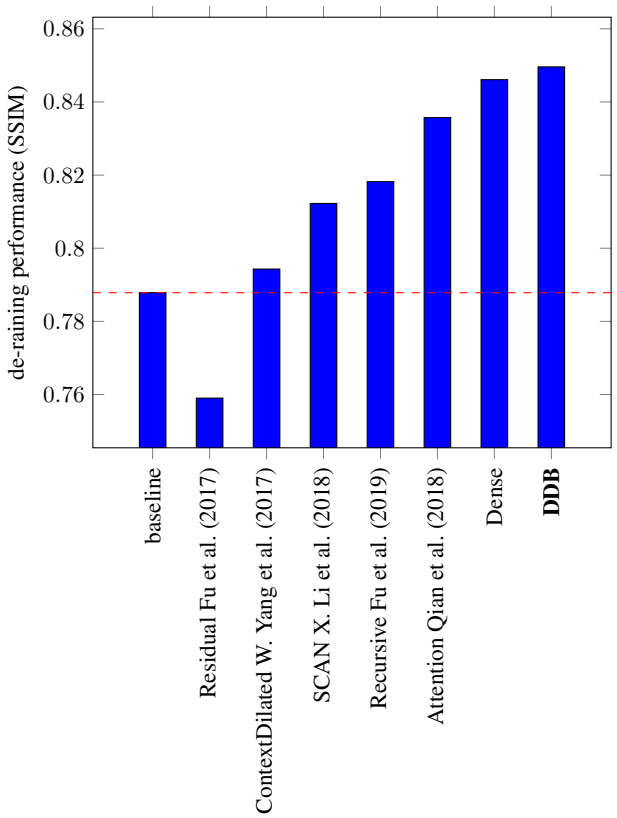

Figure 6.8: Deraining performance of models with different network modules added before the encoder of the baseline generator model.

6.7.3 Number of Network Parameters

Since the deep generative models learn the entire process of image restoration as an end-to-end mapping, they naturally require fewer numbers of network parameters than those deconstructive methods based on complicated composition models. Hence they benefit from lighter-weight models, higher flexibility, and better generalizability. In addition, the removal of excessive abstraction layers proposed in this study allows them to free more parameters from being involved in their models.

Here we compare the numbers of network parameters of an 8-layer U-Net Ronneberger et al. (2015) generator backbone used in the pixel2pixel Isola et al. (2017), a 5-layer U-Net condensed according to our experimental results on the deraining datasets, a 5-layer U-Net with a 15-layer DDB module equipped at the beginning of the encoder network, and a 5-layer U-Net with a 15-layer DDB module as well as an SCU module equipped as the upsampling operation of the top layer in the decoder network (Table 6.3).

pix2pix (UNet-8)	UNet-5	UNet-5 + DDB	UNet-5 + DDB + SCU
54,404,099	16,655,363	17,578,403	17,596,844

Table 6.3: Number of parameters compared with backbone generator network

We can see that: even though we apply additional network modules of a 15-layers DDB and an SCU module (which involve an extra convolutional layer), the new network parameters introduced by these modules do not occupy much compared with the excessive down-and-up-sampling layer removed from the backbone network. Thus, in general, our proposed methods can further reduce the number of network parameters of the models to a large extent.

6.8 Implementation Details

In this study, we modify and conduct relevant experiments mainly using the pixel2pixel model Isola et al. (2017) as our baseline method.

After applying the LSGAN loss, the overall loss function for training the GAN model are as follow:

$$\mathcal{L}_{\mathcal{G}}(\mathcal{G},\mathcal{D}) = \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[(\mathcal{D}(\mathbf{x},\mathcal{G}(\mathbf{x}))-1)^2\right] + \lambda\mathcal{L}_{L1}(\mathcal{G}) \tag{6.5}$$

$$\mathcal{L}_{\mathcal{D}}(\mathcal{G},\mathcal{D}) = \frac{1}{2}\mathbb{E}_{\mathbf{x},\mathbf{y}}\left[(\mathcal{D}(\mathbf{x},\mathbf{y})-1)^2\right] + \frac{1}{2}\mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\mathcal{D}(\mathbf{x},\mathcal{G}(\mathbf{x}))^2\right] \tag{6.6}$$

Similarly, we also include L1-loss as the complement to the discriminator on scoring low-frequency information, which can reduce blurring and guide the generator in details adjustment. In case the discriminator fails, the generator can still go in the gradient-appropriate direction.

$$\mathcal{L}_{L1}(\mathcal{G}) = \mathbb{E}_{\mathbf{x},\mathbf{y}}\left\|\mathbf{y}-\mathcal{G}(\mathbf{x})\right\|_1 \tag{6.7}$$

the L1-loss $\mathcal{L}_{L1}(\mathcal{G})$ is joined with the LSGAN MSE loss to form the generator loss (equation 6.5), with λ as a hyper-parameter.

As for the discriminator, we use a 5-layer fully convolutional network and follow the idea of PatchGAN in pixel2pixel. Since we have mentioned that an imbalance exists between the generator and the discriminator, in which the discriminator is always the first to converge and thus fails to provide the gradient to the generator to continue training. A common understanding here is that the discriminator is over-powerful than the generator. However, we also tried to reduce the number of layers and try to use some "weaker" networks as the discriminator, but all experiment ends up the same. This may illustrate that the difference between the generated data and the real ground truth does not lie on high-level features, and a shallow network can also tell their differences. Therefore, instead of elaborating the discriminator network, we try to reinforce the generator network so as to compete with the discriminator. PatchGAN here is found still useful in image restoration tasks, which processes each image patch identically and independently and guarantees that when the noise is not uniformly distributed on the input image, the discriminator can still make a general evaluation on the quality of the generated image. We also compared its performance with the multi-scale discriminator proposed in the ID-CGAN, where the experiment results turn out to be the same. So here for faster training, we do not use the multi-scale model, which includes more convolution operations.

For the training of our model, we use batch size equal to 1. For each iteration of training, image are randomly crop into a smaller size as input to the our model, in a bid to augment the training data and improve model's generalization ability. The ideal crop size are combinations among 256, 512, and 1024, which mainly depends on the datasets (the crop size should be large enough to contain as complete semantic information as possible, the minimum crop size for RainCityScapes dataset, for instance, should be 512x512). we employed Adam as the optimizer with 0.0002 as learning rate, 0.5 and 0.999 as the first and the second momentum values, and 0 as weight decay.

Relevant programmes are implemented using the platform of PyTorch and we conducted all experiments a physical environment with Intel Xeon(R) Silver 4108 as CPU and GeForce RTX 2080 Ti as GPU.

Chapter 7

Conclusion

This chapter concludes this thesis by summarizing the key contributions of this research and illustrating the limitations as well as the future works to be done.

7.1 Summary & Contributions

Image restoration tasks in physics-based vision attach great significance to the processing of visual data and the subsequent applications, where generative methods, particularly GANs, are considered promising for these tasks. Whereas, we found that existing GAN-based methods tend to be direct applications of conventional GAN models, which may fail in achieving satisfactory performances on image restoration and are less competitive to the traditional methods based on complicated handcrafted deconstructive models.

In this study, we propose an information-theoretic framework to analyze the information flows and optimization objectives of deep generative models and can help to explain the models for image restoration tasks. Based on this theory, we identify that: conventional GAN models may not be directly applicable to the tasks of image restoration and may suffer from problems of over-invested abstraction, inherent details loss, as well as gradient vanishing, and imbalance of training. Simple solutions, including optimization of network structure, enhancing details extraction, accumulation, and retention, as well as replacing the measures of the loss function can significantly improve the baseline methods, proving that the generalization ability of deep generative models like GANs is still competent to learn and simulate the complex physical integrations in the image restoration tasks.

To sum up, this study contributes in the following aspects:

- we pointed out that general GAN models for conventional generation tasks may not be directly applicable to the tasks of image restoration, and identified three key issues in the direct applications of these conventional GAN models (over-invested abstraction process, inherent details loss, as well as vanishing gradients and training imbalance) and proved with corresponding empirical evidence. This generally answered the question of why existing GAN-based methods fail to achieve satisfactory performances on image restoration tasks compared with the traditional methods using deconstructive models;

- we provide an information-theoretic framework to explain the learning processes of conventional DGMs as well as generative methods for image restoration, where we analyzed

the information flows in these models, identified the sources of information involved in generating the results, as well as elaborate on the learning objectives and optimization boundaries for different parts of the models. This can be used for the analysis of relevant methods and instruct the design of network models for image restoration or even other tasks;

- we provide the basic ideas to cope with the three identified issues above and proposed practical solutions or suggestions to implement (including the optimization of network structure, enhancing details extraction, accumulation, and retention, as well as alternating the measures of the loss function, et al.) for improving the existing GAN-based image restoration methods. This can be of great reference value for the future design of relevant network models, optimization objectives, and measures et al., and may even be generalized to other image-to-image generation tasks.

7.2 Limitations & Future Works

In this study, we only focus on the convolutional-based generator network, where information in the source inputs is sparsely processed based on their spatial neighboring on the images. The problems analyzed above may not be directly generalized to other kinds of generator networks and may need to be discussed specifically. For instance, Visual Transformer (ViT) Dosovitskiy et al. (2020) is recently proposed as a new kind of network structure that is considered promising for processing visual data. Since the ViTs may not involve the down-and-up-sampling mechanisms, problems such as over-invested abstraction may not be applicable for these kinds of network models. Thus, it can be of great research value to study the corresponding information-theoretic frameworks for the ViTs and other networks as the future works of this study.

Although we have provided corresponding solutions to the identified problems and can effectively improve the baseline models, most of these methods tend to be naive solutions and there may exist better approaches or further improvements to be applied. For example, the DDB module we propose may suffer from the redundancy of feature maps, where considerable channels that are concatenated and accumulated by the DenseNet Huang et al. (2017) structure may probably share high similarities. Model pruning or relevant post-processing may be helpful for these methods, which can also be discussed in the later works.

Moreover, the experiments and the proposed solutions are applied only based on the standard GAN model of pixel2pixel Isola et al. (2017), where more advanced network structures, loss functions, or training techniques proposed in relevant studies are not yet applied in this thesis. Theoretically, these current advances may further improve the performances of GAN models in image restoration tasks. However, since this research mainly focuses on the theoretical understanding and interpretability of the relevant models, the adoption of these ideas can be explored in further studies.

Additionally, this thesis only investigates the image deraining and dehazing performances of models trained using supervised methods, while there are many other image restoration tasks to be examined and the proposed idea may be also generalized to the unsupervised learning

processes. Future studies can be conducted to verify the theoretical framework as well as the proposed methods in different image restoration tasks, as well as their corresponding performance trained using unsupervised learning methods.

Appendix A

Proof of the Optimization Objectives & Information Boundaries

Proof. Optimization Objectives of Components & Corresponding Information Boundaries. Given the information flow and the optimization objectives of the generative deraining models as Fig. A.1:

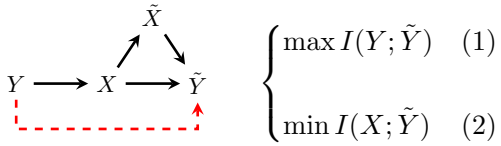

Figure A.1: Information flow and the optimization objectives (simplified).

For optimization objective (1), we have:

$$I(Y;\tilde{Y}) = I(Y;X;\tilde{X};\tilde{Y}) + I(Y;X \mid \tilde{X};\tilde{Y}) + I(Y \mid X;\tilde{Y}) \tag{A.3}$$

For the third term $I(Y \mid X;\tilde{Y})$, according to the Data Processing Inequality (DPI):

$$I(Y \mid X;\tilde{Y}) \leq I(Y \mid X) \tag{A.4}$$

For objectives (2), we have:

$$I(X;\tilde{Y}) = I(X;\tilde{X};\tilde{Y}) + I(X \mid \tilde{X};\tilde{Y}) \tag{A.5}$$

Since in the generator networks, $I(X;\tilde{X};\tilde{Y})$ are features extracted by learn-able parameters, while $I(X \mid \tilde{X};\tilde{Y})$ represents the information about X that directly passes through the generator networks without learning processes, we can have the possible ranges of the two terms above (noted that $\tilde{X}$ and $\tilde{Y}$ are the variables of optimization):

$$-H(\tilde{X}) \leq I(X;\tilde{X};\tilde{Y}) \leq H(\tilde{X}) \tag{A.6}$$

$$0 \leq I(X \mid \tilde{X};\tilde{Y}) \leq H(X) \tag{A.7}$$

Thus, as long as $H(\tilde{X}) \geq H(X \mid Y)$, we can solve the min-max problem and obtain the optimal conditions of each functional structure of the network models as follow:

$$\begin{cases} \min I(X;\tilde{X};\tilde{Y}) & I(X;\tilde{X};\tilde{Y}) \geq -H(X \mid Y) \\ \max I(X \mid \tilde{X};\tilde{Y}) & I(X \mid \tilde{X};\tilde{Y}) \leq H(X) \\ \max I(Y \mid X;\tilde{Y}) & I(Y \mid X;\tilde{Y}) \leq I(Y \mid X) \end{cases} \tag{A.8}$$

□

Bibliography

Abdelhamed, A., Lin, S., & Brown, M. S. (2018, June). A high-quality denoising dataset for smartphone cameras. In *Ieee conference on computer vision and pattern recognition (cvpr)*.

Ackermann, J., & Goesele, M. (2015). A survey of photometric stereo techniques. *Foundations and Trends® in Computer Graphics and Vision*, *9*(3-4), 149–254.

Aitken, A., Ledig, C., Theis, L., Caballero, J., Wang, Z., & Shi, W. (2017). Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize. *arXiv preprint arXiv:1707.02937*.

Alemi, A. A., Fischer, I., Dillon, J. V., & Murphy, K. (2016). Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*.

Alsaiari, A., Rustagi, R., Thomas, M. M., Forbes, A. G., et al. (2019). Image denoising using a generative adversarial network. In *2019 ieee 2nd international conference on information and computer technologies (icict)* (pp. 126–132).

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828.

Bengio, Y., Yao, L., Alain, G., & Vincent, P. (2013). Generalized denoising auto-encoders as generative models. *arXiv preprint arXiv:1305.6663*.

Bennur, A., Gaggar, M., et al. (2020). Lca-net: Light convolutional autoencoder for image dehazing. *arXiv preprint arXiv:2008.10325*.

Bhoi, A. (2019). Monocular depth estimation: A survey. *arXiv preprint arXiv:1901.09402*.

Bond-Taylor, S., Leach, A., Long, Y., & Willcocks, C. G. (2021). Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1-1. doi: 10.1109/TPAMI.2021.3116668

Bourlard, H., & Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, *59*(4), 291–294.

Brand, M. (1997). Physics-based visual understanding. *Computer Vision and Image Understanding*, *65*(2), 192–205.

Brooks, A. C., Zhao, X., & Pappas, T. N. (2008). Structural similarity quality metrics in a coding context: exploring the space of realistic distortions. *IEEE Transactions on image processing*, *17*(8), 1261–1273.

Buhrmester, V., Münch, D., & Arens, M. (2021). Analysis of explainers of black box deep neural networks for computer vision: A survey. *Machine Learning and Knowledge Extraction*, *3*(4), 966–989.

Cai, B., Xu, X., Jia, K., Qing, C., & Tao, D. (2016). Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, *25*(11), 5187–5198.

Chen, J., Chen, J., Chao, H., & Yang, M. (2018). Image blind denoising with generative adversarial network based noise modeling. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3155–3164).

Chen, R., & Lai, E. M.-K. (2019). Convolutional autoencoder for single image dehazing. In *Icip* (pp. 4464–4468).

Chen, S., Shi, D., Sadiq, M., & Cheng, X. (2020). Image denoising with generative adversarial

networks and its application to cell image enhancement. *IEEE Access*, *8*, 82819–82831.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, *29*.

Chengtao, C., Qiuyu, Z., & Yanhua, L. (2015). A survey of image dehazing approaches. In *The 27th chinese control and decision conference (2015 ccdc)* (pp. 3964–3969).

Cover Thomas, M., & Thomas Joy, A. (1991). Elements of information theory. *New York: Wiley*, *3*, 37–38.

Csáji, B. C., et al. (2001). Approximation with artificial neural networks. *Faculty of Sciences, Etvs Lornd University, Hungary*, *24*(48), 7.

Dargan, S., Kumar, M., Ayyagari, M. R., & Kumar, G. (2019). A survey of deep learning and its applications: A new paradigm to machine learning. *Archives of Computational Methods in Engineering*, 1–22.

Deng, L.-J., Huang, T.-Z., Zhao, X.-L., & Jiang, T.-X. (2018). A directional global sparse model for single image rain removal. *Applied Mathematical Modelling*, *59*, 662–679.

Deng, S., Wei, M., Wang, J., Liang, L., Xie, H., & Wang, M. (2019). Drd-net: Detail-recovery image deraining via context aggregation networks. *arXiv preprint arXiv:1908.10267*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, Y., Xu, J., Qiu, Q., Zhen, X., & Zhang, L. (2020). Variational image deraining. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 2406–2415).

Du, Y., Xu, J., Zhen, X., Cheng, M.-M., & Shao, L. (2020). Conditional variational image deraining. *IEEE Transactions on Image Processing*, *29*, 6288–6301.

Elharrouss, O., Almaadeed, N., Al-Maadeed, S., & Akbari, Y. (2020). Image inpainting: A review. *Neural Processing Letters*, *51*(2), 2007–2028.

Emami, H., Aliabadi, M. M., Dong, M., & Chinnam, R. (2020). Spa-gan: Spatial attention gan for image-to-image translation. *IEEE Transactions on Multimedia*.

Engin, D., Genç, A., & Kemal Ekenel, H. (2018). Cycle-dehaze: Enhanced cyclegan for single image dehazing. In *Proceedings of the ieee conference on computer vision and pattern recognition workshops* (pp. 825–833).

Fan, L., Zhang, F., Fan, H., & Zhang, C. (2019). Brief review of image denoising techniques. *Visual Computing for Industry, Biomedicine, and Art*, *2*(1), 1–12.

Farsiu, S., Robinson, D., Elad, M., & Milanfar, P. (2004). Advances and challenges in super-resolution. *International Journal of Imaging Systems and Technology*, *14*(2), 47–57.

Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., & Paisley, J. (2017). Removing rain from single images via a deep detail network. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3855–3863).

Fu, X., Liang, B., Huang, Y., Ding, X., & Paisley, J. (2019). Lightweight pyramid networks for image deraining. *IEEE transactions on neural networks and learning systems*, *31*(6), 1794–1807.

Gao, L., Song, J., Liu, X., Shao, J., Liu, J., & Shao, J. (2017). Learning in high-dimensional multimedia data: the state of the art. *Multimedia Systems*, *23*(3), 303–313.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Martinez-Gonzalez, P., & Garcia-Rodriguez, J. (2018). A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing*, *70*, 41–65.

Gheisari, M., Wang, G., & Bhuiyan, M. Z. A. (2017). A survey on deep learning in big data. In *2017 ieee international conference on computational science and engineering (cse) and ieee international conference on embedded and ubiquitous computing (euc)* (Vol. 2, pp.

173–180).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* (Vol. 1) (No. 2).

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*, *27*.

Goyal, B., Dogra, A., Agrawal, S., Sohi, B. S., & Sharma, A. (2020). Image denoising review: From classical to state-of-the-art approaches. *Information fusion*, *55*, 220–244.

Gui, J., Cong, X., Cao, Y., Ren, W., Zhang, J., Zhang, J., & Tao, D. (2021). A comprehensive survey on image dehazing based on deep learning. *arXiv preprint arXiv:2106.03323*.

Gunturk, B., & Li, X. (2018). *Image restoration*. CRC Press.

Habibi, A. (1972). Two-dimensional bayesian estimate of images. *Proceedings of the IEEE*, *60*(7), 878–883.

He, K., Sun, J., & Tang, X. (2010). Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, *33*(12), 2341–2353.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation*, *14*(8), 1771–1800.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, *313*(5786), 504–507.

Hinton, G. E., Sejnowski, T. J., et al. (1986). Learning and relearning in boltzmann machines. *Parallel distributed processing: Explorations in the microstructure of cognition*, *1*(282-317), 2.

Hong, Z., Fan, X., Jiang, T., & Feng, J. (2020). End-to-end unpaired image denoising with conditional adversarial networks. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, pp. 4140–4149).

Hordley, S. D. (2006). Scene illuminant estimation: past, present, and future. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, *31*(4), 303–314.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, *2*(5), 359–366.

Horé, A., & Ziou, D. (2010). Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition* (p. 2366-2369). doi: 10.1109/ICPR.2010.579

Hu, X., Fu, C.-W., Zhu, L., & Heng, P.-A. (2019). Depth-attentional features for single-image rain removal. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 8022–8031).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4700–4708).

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1125–1134).

Jain, V., & Seung, S. (2008). Natural image denoising with convolutional networks. *Advances in neural information processing systems*, *21*.

Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., & Bengio, Y. (2017). The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Proceedings of the ieee conference on computer vision and pattern recognition workshops* (pp. 11–19).

Jeon, I., Lee, W., Pyeon, M., & Kim, G. (2021). Ib-gan: Disengangled representation learning with information bottleneck generative adversarial networks..

Kang, L.-W., Lin, C.-W., & Fu, Y.-H. (2011). Automatic single-image-based rain streaks removal via image decomposition. *IEEE transactions on image processing*, *21*(4), 1742–1755.

Karanam, S. R., Srinivas, Y., & Krishna, M. V. (2020). Study on image processing using deep learning techniques. *Materials Today: Proceedings*.

Kim, H.-J., & Lee, D. (2020). Image denoising with conditional generative adversarial networks (cgan) in low dose chest images. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *954*, 161914.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Koppal, S. J. (2020). Lambertian reflectance. *Computer vision: a reference guide*, 1–3.

Li, B., Peng, X., Wang, Z., Xu, J., & Feng, D. (2017). Aod-net: All-in-one dehazing network. In *Proceedings of the ieee international conference on computer vision* (pp. 4770–4778).

Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., & Wang, Z. (2018). Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, *28*(1), 492–505.

Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., & Wang, Z. (2019). Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, *28*(1), 492–505.

Li, R., Cheong, L.-F., & Tan, R. T. (2019). Heavy rain image restoration: Integrating physics model and conditional adversarial learning. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1633–1642).

Li, R., Pan, J., Li, Z., & Tang, J. (2018). Single image dehazing via conditional generative adversarial network. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 8202–8211).

Li, S., Araujo, I. B., Ren, W., Wang, Z., Tokuda, E. K., Junior, R. H., ... Cao, X. (2019). Single image deraining: A comprehensive benchmark analysis. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 3838–3847).

Li, X., Wu, J., Lin, Z., Liu, H., & Zha, H. (2018). Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *Proceedings of the european conference on computer vision (eccv)* (pp. 254–269).

Li, Y., Fu, Y., Liang, S., & Zheng, Y. (2019). *Physics based vision meets deep learning.* The 2nd International Workshop on Physics Based Vision meets Deep Learning (PBDL). Retrieved from `https://pbdl2019.github.io/challenge/index.html` ([Online])

Li, Y., Tan, R. T., Guo, X., Lu, J., & Brown, M. S. (2016). Rain streak removal using layer priors. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2736–2744).

Liu, J., Yang, W., Yang, S., & Guo, Z. (2018). Erase or fill? deep joint recurrent rain removal and reconstruction in videos. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3233–3242).

Luo, Y., Xu, Y., & Ji, H. (2015). Removing rain from a single image via discriminative sparse coding. In *Proceedings of the ieee international conference on computer vision* (pp. 3397–3405).

Majumdar, A. (2018). Blind denoising autoencoder. *IEEE transactions on neural networks and learning systems*, *30*(1), 312–317.

Malav, R., Kim, A., Sahoo, S. R., & Pandey, G. (2018). Dhsgan: An end to end dehazing network for fog and smoke. In *Asian conference on computer vision* (pp. 593–608).

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., & Paul Smolley, S. (2017). Least squares

generative adversarial networks. In *Proceedings of the ieee international conference on computer vision* (pp. 2794–2802).

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information, henry holt and co. *Inc., New York, NY*, *2*(4.2).

McCartney, E. J. (1976). Optics of the atmosphere: scattering by molecules and particles. *New York*.

Metaxas, D. N. (2012). *Physics-based deformable models: applications to computer vision, graphics and medical imaging* (Vol. 389). Springer Science & Business Media.

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Motwani, M. C., Gadiya, M. C., Motwani, R. C., & Harris, F. C. (2004). Survey of image denoising techniques. In *Proceedings of gspx* (Vol. 27, pp. 27–30).

Nahi, N. E. (1972). Role of recursive estimation in statistical image enhancement. *Proceedings of the IEEE*, *60*(7), 872–877.

Narasimhan, S. G., & Nayar, S. K. (2000). Chromatic framework for vision in bad weather. In *Proceedings ieee conference on computer vision and pattern recognition. cvpr 2000 (cat. no. pr00662)* (Vol. 1, pp. 598–605).

Nasrollahi, K., & Moeslund, T. B. (2014). Super-resolution: a comprehensive survey. *Machine vision and applications*, *25*(6), 1423–1468.

Nixon, M., & Aguado, A. (2019). *Feature extraction and image processing for computer vision*. Academic press.

Noh, H., Hong, S., & Han, B. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the ieee international conference on computer vision* (pp. 1520–1528).

Oussidi, A., & Elhassouny, A. (2018). Deep generative models: Survey. In *2018 international conference on intelligent systems and computer vision (iscv)* (pp. 1–8).

Pan, J., Dong, J., Liu, Y., Zhang, J., Ren, J., Tang, J., ... Yang, M.-H. (2020). Physics-based generative adversarial models for image restoration and beyond. *IEEE transactions on pattern analysis and machine intelligence*, *43*(7), 2449–2462.

Parker, J. R. (2010). *Algorithms for image processing and computer vision*. John Wiley & Sons.

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., ... Iyengar, S. (2018). A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR)*, *51*(5), 1–36.

Qian, R., Tan, R. T., Yang, W., Su, J., & Liu, J. (2018). Attentive generative adversarial network for raindrop removal from a single image. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2482–2491).

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

Ranzato, M., Poultney, C., Chopra, S., LeCun, Y., et al. (2007). Efficient learning of sparse representations with an energy-based model. *Advances in neural information processing systems*, *19*, 1137.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241).

Ruthotto, L., & Haber, E. (2021). An introduction to deep generative modeling. *GAMM-Mitteilungen*, *44*(2), e202100008.

Salakhutdinov, R., Mnih, A., & Hinton, G. (2007). Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on machine learning* (pp. 791–798).

Shafer, S. A., Kanade, T., Klinker, G. J., & Novak, C. L. (1990). Physics-based models for early vision by machine. In *Perceiving, measuring, and using color* (Vol. 1250, pp. 222–235).

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, *5*(1), 3–55.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., ... Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1874–1883).

Smolensky, P. (1986). *Information processing in dynamical systems: Foundations of harmony theory* (Tech. Rep.). Colorado Univ at Boulder Dept of Computer Science.

Sun, Y., Liu, X., Cong, P., Li, L., & Zhao, Z. (2018). Digital radiography image denoising using a generative adversarial network. *Journal of X-ray Science and Technology*, *26*(4), 523–534.

Szeliski, R. (2010). *Computer vision: algorithms and applications*. Springer Science & Business Media.

Tan, R. T. (2008). Visibility in bad weather from a single image. In *2008 ieee conference on computer vision and pattern recognition* (pp. 1–8).

Tekalp, A. M., et al. (2022). Deep learning for image/video restoration and super-resolution. *Foundations and Trends® in Computer Graphics and Vision*, *13*(1), 1–110.

Tian, C., Fei, L., Zheng, W., Xu, Y., Zuo, W., & Lin, C.-W. (2020). Deep learning on image denoising: An overview. *Neural Networks*, *131*, 251–275.

Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. *arXiv preprint physics/0004057*.

Tishby, N., & Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)* (pp. 1–5).

Vastola, K., & Poor, H. (1984). Robust wiener-kolmogorov theory. *IEEE Transactions on Information Theory*, *30*(2), 316–327.

Wan, R., Shi, B., Duan, L.-Y., Tan, A.-H., & Kot, A. C. (2017). Benchmarking single-image reflection removal algorithms. In *Proceedings of the ieee international conference on computer vision* (pp. 3922–3930).

Wolff, L. B., Shafer, S. A., & Healey, G. E. (1993). *Physics-based vision: Principles and practice: Shape recovery, volume 3* (Vol. 3). CRC Press.

Wyatt, C. (2012). *Radiometric calibration: theory and methods*. Elsevier.

Xiong, F., Wang, Q., & Gao, Q. (2019). Consistent embedded gan for image-to-image translation. *IEEE Access*, *7*, 126651–126661.

Yang, Q., Yan, P., Zhang, Y., Yu, H., Shi, Y., Mou, X., ... Wang, G. (2018). Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE transactions on medical imaging*, *37*(6), 1348–1357.

Yang, W., Tan, R. T., Feng, J., Liu, J., Guo, Z., & Yan, S. (2017). Deep joint rain detection and removal from a single image. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1357–1366).

Yang, W., Tan, R. T., Wang, S., Fang, Y., & Liu, J. (2020a). Single image deraining: From model-based to data-driven and beyond. *IEEE Transactions on pattern analysis and machine intelligence*, *43*(11), 4059–4077.

Yang, W., Tan, R. T., Wang, S., Fang, Y., & Liu, J. (2020b). Single image deraining: From model-based to data-driven and beyond. *IEEE Transactions on pattern analysis and machine intelligence*.

Yeh, R. A., Lim, T. Y., Chen, C., Schwing, A. G., Hasegawa-Johnson, M., & Do, M. N. (2018). Image restoration with deep generative models. In *2018 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 6772–6776).

Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Yu, W., Huang, Z., Zhang, W., Feng, L., & Xiao, N. (2019). Gradual network for single image de-raining. In *Proceedings of the 27th acm international conference on multimedia* (pp. 1795–1804).

Yu, Y. (2021). *Physics-based vision meets deep learning* (Unpublished doctoral dissertation). University of York.

Zhang, H., & Patel, V. M. (2018). Density-aware single image de-raining using a multi-stream dense network. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 695–704).

Zhang, H., Sindagi, V., & Patel, V. M. (2019). Image de-raining using a conditional generative adversarial network. *IEEE transactions on circuits and systems for video technology*.

Zhang, K., Zuo, W., Chen, Y., Meng, D., & Zhang, L. (2017). Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, *26*(7), 3142–3155.

Zhang, R., Tsai, P.-S., Cryer, J. E., & Shah, M. (1999). Shape-from-shading: a survey. *IEEE transactions on pattern analysis and machine intelligence*, *21*(8), 690–706.

Zhao, Z. (2012). Image denoising by autoencoder: Learning core representations. *The Australian National University*.

Zhao, Z., Zheng, P., Xu, S., & Wu, X. (2019). Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, *30*(11), 3212–3232.

Zhong, G., Wang, L.-N., Ling, X., & Dong, J. (2016). An overview on data representation learning: From traditional feature learning to recent deep learning. *The Journal of Finance and Data Science*, *2*(4), 265–278.

Zhou, Y., Chellappa, R., & Jenkins, B. (1987). A novel approach to image restoration based on a neural network. In *Proceedings of the international conference on neural networks, san diego, california.*

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the ieee international conference on computer vision* (pp. 2223–2232).

Zhu, J.-Y., Zhang, R., Pathak, D., Darrell, T., Efros, A. A., Wang, O., & Shechtman, E. (2017). Toward multimodal image-to-image translation. In *Advances in neural information processing systems* (pp. 465–476).

Zhu, L., Fu, C.-W., Lischinski, D., & Heng, P.-A. (2017). Joint bi-layer optimization for single-image rain streak removal. In *Proceedings of the ieee international conference on computer vision* (pp. 2526–2534).