UNIVERSITY OF CALIFORNIA
Los Angeles

Using AI to Mitigate Variability in CT Scans: Improving Consistency in Medical Image Analysis

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Electrical and Computer Engineering

by

Leihao Wei

2021

ABSTRACT OF THE DISSERTATION

Using AI to Mitigate Variability in CT Scans: Improving Consistency in Medical Image Analysis

by

Leihao Wei
Doctor of Philosophy in Electrical and Computer Engineering
University of California, Los Angeles, 2021
Professor Gregory J Pottie, Co-Chair
Professor William Hsu, Co-Chair

Computed tomography (CT) plays an integral role in diagnosing and screening various types of diseases. A growing number of machine learning (ML) models have been developed for prediction and classification that utilize derived quantitative image features, thanks in part to the availability of large CT datasets and advances in medical image analysis. Researchers have classified disease severity using quantitative image features such as hand-crafted radiomic and deep features. Despite reporting high classification performance, these models typically do not generalize well. Variations in the appearance of CT scans caused by differences in acquisition and reconstruction parameters adversely impact the reproducibility of quantitative image features and the performance of machine learning algorithms. As a result, few ML algorithms have been used in clinical settings. Mitigating the effects of varying CT acquisition and reconstruction parameters is a challenging inverse problem. Recent advances in deep learning have demonstrated that image translation and denoising models can achieve high per-pixel similarity metrics when compared to a target image. The purpose of this dissertation is to develop and evaluate two conditional generative models that mitigate the effects of working with CT scans acquired and reconstructed with a variety of parameters. The overarching hypothesis is that improved image quality results in better consistency in nodule detection. In essence, these models attempt to learn the underlying conditional distribution on the normalized images (high-quality) given the un-normalized (low-quality) images. First, I propose a novel CT image normalization method based on a 3D conditional generative adversarial network (GAN) that utilizes a spectral-normalization algorithm. My model provides an end-to-end solution for normalizing scans acquired using different doses, slice thicknesses, and reconstruction kernels. This study demonstrates that the GAN is capable of mitigating the variability in image quality, quantitative image features, and lung nodule detection using an automated Computer-Aided-Detection (CAD) algorithm. I show that GAN improved perceptual similarity by 22%, and resulted in a 16% increase in features with a good level of agreement based on concordance correlation coefficient analysis. As a result, the performance of the existing nodule detection model was up to 75% more consistent with the reference scan. Second, I explore the use of a conditional normalizing flow-based model to incorporate uncertainty information during image translation. The model is capable of learning the explicit conditional density and generating several plausible image outputs, providing a means to reduce the distortions introduced by existing methods. I show that the normalizing flow method achieves a 6% improvement in perpetual quality compared to the state-of-the-art GAN-based method and the resulted agreement level of the detection task is improved by 13%. This dissertation compares these two generative approaches, identifying their strengths and limitations when normalizing heterogeneous CT images and mitigating the effect of different acquisition and reconstruction parameters on downstream clinical tasks.

The dissertation of Leihao Wei is approved.

Jonathan Kao

Jason Cong

William Hsu, Committee Co-Chair

Gregory J Pottie, Committee Co-Chair

University of California, Los Angeles

2021

The dissertation is dedicated to my parents and to my wife.

Vires in Numeris

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my gratitude and respect for my advisor, Dr. William Hsu, for his trust in me to join his research group. Thank you for your assistance in helping me to advance in both my personal and professional development. Under his guidance, I learned to become a better independent researcher to tackle the challenges of AI in medical imaging. He proved to be a great mentor to me when I was battling against my medical conditions while writing this dissertation. I appreciate his patience immensely. As an expert in this field, he showed great vision to lead the project, taught me so much to develop my technical skills, and offered me resources to expand this research work. Thanks to Alex Bui, director of MII, for assembling this incredible team, supporting our daily operations, and most importantly, introducing me to Dr. Hsu and providing me with funding and opportunities to complete my PhD research. I want to thank my committee chair, Dr. Gregory Pottie, who was kind enough to let me remain under his supervision at the Electrical and Computer Engineering Department. Each time we meet, I learn new research concepts. I want also to thank Drs. Jason Cong and Jonathan Kao for serving as my committee members to provide valuable feedback on the dissertation. I am thankful to Drs. Ashely Prosper and Denise Aberle for their domain knowledge in lung cancer practice that drives each iteration of the AI model development. Thanks to our collaborators, Mike McNitt-Gray, Matt Brown, Nastaran Emaminejad, and M. Wasil Wahi-Anwar, for their generosity to share us with UCLA's dataset and thoughtful discussions regarding GAN normalization research and paper revisions.

My greatest gratitude goes to the love of my life, for her love, support, and sacrifices during my PhD program. PhDs are admittedly difficult to obtain. However, she made my life delightful. Without her assistance, this dissertation would not have been completed.

I want to thank Drs. Jennifer Casey, Jonny Pang, and the Department of Chemistry for providing me with the Teaching Assistant Fellowship during my transition to do research at MII.

Specifically, I would like to thank Denise Luna, Isabel Rippy, and Patrick Langdon for their timely assistance in setting up the lab computers and dealing with administrative matters. Thank you to all lovely current and past lab members at MII who offered my necessary help, academically or personally: Yannan Lin, Panayiotis Petousis, Nova Smedley, Rina Ding, Simon Han, Akinyinka Omigbodun, Noor Nakhaei, Anil Yadav, Chrysostomos Marasinou, Tianqing Li, Jiayun Li, Wenyuan Li, Karthik Sarma, Zichen Wang, Jennifer Polson, Haoyue Zhang, Yiwen Meng, and Tianran Zhang.

Last but not least, I would like to express my gratitude to my friends, colleagues, and those whom I have not yet mentioned. In the best six years of my life, they have brought me joy, shared my pain, and grown in wisdom together.

VITA

2010–2014 B.S. in Electrical and Computer Engineering, Rose-Hulman Institute of Technology, Terre Haute, USA.

2014–2015 M.S. in Electrical and Computer Engineering, Rose-Hulman Institute of Technology, Terre Haute, USA.

2015-2021 Graduate Student Researcher at Electrical and Computer Engineering, UCLA, Los Angeles, USA.

2018 Research intern at Meta Co.

2020 Software engineer intern at Facebook Inc.

PUBLICATIONS

L. Wei, Y. Lin, and W. Hsu, "Using a Generative Adversarial Network for CT Normalization and its Impact on Radiomic Features," in 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 844–848, IEEE, 2020

Y. Lin, L. Wei, S. X. Han, D. R. Aberle, and W. Hsu, "Edicnet: An End-to-end Detection and Interpretable Malignancy Classification Network for Pulmonary Nodules in Computed Tomography," in Medical Imaging 2020: Computer-Aided Diagnosis, vol. 11314, p. 113141H, International Society for Optics and Photonics, 2020

L. Wei and W. Hsu, "Efficient and Accurate Spatial-temporal Denoising Network for Low-dose CT scans," in Medical Imaging with Deep Learning (MIDL), 2021

T. Li, L. Wei, and W. Hsu, "A Multi-pronged Evaluation for Image Normalization Techniques," in 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 1292–1296, IEEE, 2021

L. Wei, N. Emaminejad, M. W. Wahi-Anwar, M. F. McNitt-Gray, M. S. Brown, and W. Hsu, "Using a Spectral-norm Generative Adversarial Network to Mitigate Variability in Computed Tomography Scans (in review),"Journal of Biomedical and Health Informatics, 2021

L. Wei, A. Yadav, T. Li, and W. Hsu, "Mitigating Effects of Variations in Computed Tomography Using Normalizing Flow (in review)," Transactions on Medical Imaging, 2021

CHAPTER 1

Introduction

Computer vision entered the era of deep learning since the introduction of AlexNet [58] in 2012, which won the 2012 ImageNet LSVRC-2012 competition by a large margin compared to second place (15.3% vs 26.2% error rates). The performance of machine learning (ML)-based object detection consistently improved over the years. Object detection is not the only application of deep learning. It can also be used in imaging applications for segmentation [77], image generation [36], image restoration [121], and non-imaging applications such as natural language processing (NLP) [89]. This remarkable achievement in artificial intelligence (AI) would not have been possible without the modern data infrastructure that enables us to process, store, and transfer large amounts of data at a lower cost. Together, with the use of Graphic Processing Units (GPUs) for model training, these advances have led to a new stage in which professions, particularly medicine, can employ easy-to-access tools to harness the power of image analysis powered by deep learning models to help make better decisions (e.g., become better at identifying subtle signs of early-stage disease).

Clinical decisions based solely on AI output can have life-or-death consequences. The FDA has set strict requirements for medical device licensing due to the high risk involved in applying AI/ML in medical diagnosis and decision-making. As of September 2020, 29 medical devices refer to AI/ML have been registered with the Food & Drug Administration [12]. Though AI/ML-based medical solutions are increasingly available on the market, adoption remains a challenge. The implementation of these technologies in medical practice is hindered by regulatory frameworks and a lack of trust between physicians and patients regarding new technologies. Researchers in the field of AI are responsible for bridging the knowledge gap by ensuring that models are reliable and reproducible with the goal of achieving more consistent clinical outcomes.

This dissertation demonstrates how AI can be harnessed to enhance diagnostic images, making them more consistent with the goal of improving clinical outcomes. While these techniques can be applied in a variety of clinical domains and imaging modalities, this work is driven by the a desire to improve the detection of early-stage lung cancer from Computed Tomography (CT) scans.

1.1 Motivation

The World Health Organization estimates that lung cancer will cause 1.8 million deaths in 2020 [114]. Lung cancer is the leading cause of death compared to cancers of the colon, breast, and prostate combined. As a result of lower smoking rates and improvements in early detection and treatment, lung cancer cases are continuously declining. CT is widely used to screen for lung cancer largely due to its high sensitivity, detecting even very small nodules in the lung. Low-Dose Computed Tomography (LDCT) of the chest is particularly effective and safe in detecting lung cancer at its earliest, most treatable stage. As a randomized controlled trial, the National Lung Screening Trial (NLST) [120] showed a 20% mortality rate reduction in patients who underwent chest low-dose computed tomography (LDCT). The key to improving the survival rate and prognosis of cancer patients lies in early detection and intervention.

The increasing availability of large CT datasets and advances in medical image analysis have resulted in a proliferation of machine learning (ML) models that use images for classification and prediction. Researchers from Google Artificial Intelligence and Northwestern Medicine have developed an AI model capable of detecting lung cancer more accurately than radiologists [8]. However, despite its success, the algorithm has not been implemented in clinical practice in the two years since publication. Jacob and Ginneken [50] noted that although the model is promising, further validation is required, and it could only be used if the Lung-RADS [83] screening guidelines were changed to allow for recommendations from proprietary AI systems.

In addition, both the model training and validation were conducted using NLST data. Whether the model could handle heterogeneous datasets effectively is uncertain. A large body of research has demonstrated that quantitative imaging features (QIFs), including hand-crafted "radiomic" features and neural network-based "deep" features, can be used to predict disease severity and progression. However, CT scan acquisition and reconstruction variants have a significant impact on the outcome, making QIFs with poor reproducibility. Variations in dose, slice thickness, reconstruction method, and reconstruction kernel negatively impact the reproducibility of QIFs and the performance of ML models that detect lung nodules [28–30]. To date, solutions have been to standardize acquisition protocols prospectively, which excludes the analysis of existing scans, or to normalize post-reconstructed images, which have had mixed results. Moreover, determining the optimal strategy for image normalization is task-dependent. The goal of lung cancer detection, for example, is to identify small areas of high contrast changes that could indicate suspicious nodules.

The performance of detection is affected by the dose and reconstruction method. The impact of CT parameters is not uniform, and no single approach to image normalization is optimal for all CT parameters and tasks. In this dissertation, we test the hypothesis that a systematic, task-dependent methodology for characterizing and mitigating the impact of variability on CT parameters will identify reproducible QIFs and lead to more consistent ML models.

Considering the widespread use of computer vision combined with deep learning for natural images, sometimes we may overlook the importance of AI-driven image enhancement technology. The quality of an image captured by a smartphone camera is now comparable to that captured by a digital single-lens reflex (DSLR) camera. The reason is not technological advances in optical components but computational photography powered by algorithms. In photography, this paradigm shift has already taken place. The medical imaging society can adopt this methodology as a result of the momentum of this transformation.

Ultimately, the project will use AI to improve image quality "computationally" to achieve consistency in image style and appearance, and downstream image analysis tasks can be automated without any human involvement to exclude existing scans that do not conform to the standardized protocols. This study develops robust ML models capable of mitigating the variability of medical images under various conditions simultaneously to deliver more consistent clinical predictions after normalization. However, it is challenging to achieve this goal. The appearance and use-case of medical images differ significantly from natural images. In particular, the differences impose two obstacles that prevent us from developing practical models.

- In contrast to CT scanners, cameras sensors are very standardized, ensuring relative consistency in image quality across a range of devices. Technologists, on the other hand, adjust CT scanner acquisition and reconstruction parameters according various factors, including patient characteristics, institutional standards, and manufacturer recommendations. All medical imaging techniques are subject to inter- and intra-variability. As such, reconstructed images are not usually consistent in appearance.
- Due to patient privacy regulations and the low incidence of medical imaging scans, imaging data are not as readily available or as abundant as natural images. Limited data makes it difficult to train effective models. Models do not usually generalize well to data obtained from another institution.

Deep learning has been increasingly used to manipulate medical images, including low-dose CT denoising [17] and CT image synthesis [142]. As a result, recent advances in machine learning-based image restoration and super-resolution developed for natural images can also be applied to medical images.

1.2 Contribution

This dissertation addresses these two obstacles by pursuing three research aims.

- Aim 1: develop a generative adversarial network (GAN)-based model to mitigate the image feature variability under different CT image conditions.
- Aim 2: investigate an alternative normalizing Flow-based approach with the goal of improving model performance, minimizing artifacts during mitigation, and increasing trust in the model's output by estimating model uncertainty.
- Aim 3: improve the robustness of the models developed in Aims 1 & 2 by 1) improving model training and inference efficiency; 2) understanding and investigating barriers that result in poor model generalizability.

Towards Aim 1, we propose a 3D spectral-norm GAN (SNGAN) that normalizes CT images acquired and reconstructed under different conditions. We used raw projection data from LDCT chest scans of patients to generate a variety of reconstructions that simulate different dose levels, slice thicknesses, and reconstruction kernels. Each scan from 186 patients was reconstructed using 10 image conditions representing different acquisition and reconstruction parameters. Defining one condition to be the reference (i.e., a scan acquired at 100% dose, 1.0 mm slice thickness, medium kernel), we trained SNGAN models to normalize all other scans to the reference. We evaluated SNGAN against other state-of-the-art methods by comparing image quality metrics, impact on computed radiomic feature values, and a specific task (i.e., lung nodule detection). Our SNGAN improved perceptual similarity by 22%, compared to another GAN-based method. SNGAN resulted in smaller radiomic feature errors when compared to the reference condition (16% increase in features with "good" and "moderate" agreement based on concordance correlation coefficient). Performance of the existing nodule detection model was more consistent on scans normalized using SNGAN achieved compared to unnormalized scans (up to 75% improvement in concordance correlation coefficient). Collectively, these results demonstrate the SNGAN's ability to normalize heterogeneous CT images and reduce adverse impacts on downstream tasks.

Towards Aim 2, we present CTFlow, a normalizing Flow-based method for translating images acquired and reconstructed using different doses and kernels to a reference scan. Unlike existing state-of-the-art image denoising and translation approaches that only generate a single output, Flow-based methods learn the explicit conditional density, capture the uncertainty associated with restoration, and output the entire spectrum of plausible solutions. We harness these capabilities to generate more realistic restored reference scans. To evaluate the performance of CTFlow, first, we compare CTFlow with other denoising techniques by training and testing it on the AAPM-Mayo Clinic Low-Dose CT Grand Challenge dataset. CTFlow achieves superior performance for both peak signal-to-noise ratio and perceptual quality metrics. Second, we train and evaluate CTFlow on the same CT chest scans collected in Aim 1, analyzing the difference in restored reference scans on the performance of a lung nodule detection algorithm. CTFlow produces more consistent predictions across all dose and kernel conditions than the GAN-based method in Aim 1. Third, we investigated generalization performance by evaluating a pretrained CTFlow model on a publicly available low-dose CT chest dataset. We show that CTFlow maintains higher image fidelity than GAN-based methods. In summary, normalizing flow performs state-of-the-art CT image translation and provides additional information through its ability to quantify restoration uncertainty.

Towards Aim 3, as an extension of both Aim 1 and 2, we performed two analyses. First, we present an efficient and accurate spatial-temporal convolution method to accelerate an existing denoising network based on the SRResNet. We trained and evaluated our model using data from our institution. We compared the performance of the proposed spatial-temporal convolution network to the SRResNet with full 3D convolutional layers. Using 8-bit quantization, we demonstrated a 7-fold speed-up during inference. Using lung nodule characterization as a driving task, we analyzed the impact on image quality metrics and radiomic feature values. Our results show that our method achieves better perceptual quality, and the outputs are consistent with the SRResNet baseline outputs for some radiomic features (31 out of 57 total features). These observations together demonstrate that the proposed spatial-temporal method can be potentially useful for clinical applications where the computational resource is limited.

Second, we evaluated image normalization techniques using a multi-pronged approach that incorporates 1) per-pixel image quality, 2) radiomic features variability, and 3) task performance differences, using a ML model. As part of the evaluation of a previously reported 3D GAN-based approach in Aim 1, we examined its performance on LDCT scans acquired at different institutions with varying dose levels and reconstruction kernels. However, the GAN did not improve performance in terms of quantitative imaging features or downstream tasks even though it produced superior image quality metrics. In summary, these results suggest a more complex relationship between CT acquisition and reconstruction parameters and their effect on radiomic features and ML model performance, which cannot be captured by using pixel metrics alone. Our approach provides a more comprehensive picture of the effect of normalization. Both efficiency and generalization studies are under the umbrella of model robustness.

Collectively, this dissertation provides the medical imaging community with a ML framework for image normalization and touches upon topics such as image synthesis, model generalization, task-driven evaluations, and model uncertainty analysis. Under this framework, results from this work provide evidence that ML-based methods are capable of normalizing heterogeneous scans acquired under different conditions to a reference condition (e.g., parameters that are recommended by Societies for clinical practice). This work helps address part of the gap between computer-aided detection/diagnosis models that are cited in the academic literature and their clinical translation.

1.3 Organization

In Chapter 2, we describe the technical background information on the basics of CT, radiomics, and an overview of related work of image enhancement techniques for mitigating variability in medical images. Chapter 3 discusses Aim 1, developing and evaluating a GAN to normalize CT images. This chapter is an extension work based on a previous peer-reviewed paper [133]. In order to overcome the limitations of the model described in Chapter 3, Chapter 4 is related to Aim 2, describing the conditional Flow-based image normalization model. I show how this approach is capable of modeling uncertainty which is unsupported by GAN-based methods. Chapter 5 summarizes the comprehensive evaluation performed as part of Aim 3 and proposes recommendations for improving model generalization. This chapter is adapted from a prior publication [69]. Moreover, Chapter 5 discusses ways to improve computational efficiency using the novel spatial-temporal convolution. A summary of the results and concluding remarks from Chapters 3 to 5 are presented in Chapter 6, along with a discussion of the limitations of this work and future directions.

CHAPTER 2

Background

2.1 Computed Tomography

A beam of x-rays is aimed at the patient at different angles, creating a series of signals that are analyzed by the machine's computer to produce cross-sectional images of the body. Data from the scanner were transmitted to a computer, which combined successive slices into a three-dimensional image of the patient, making it easier to locate and identify basic organs, tissues, and tumors. The process to map the measurements back to a three-dimensional image, is considered an inverse problem. In this problem setting, $\mathcal{M}$ is an operator that yields measurement data y given the model parameters x, the true image data, and $\mathcal{N}$ is the measurement noise. The goal of CT imaging is to infer the true image from the measurement results y.

$$y = \mathcal{M}(x) + \mathcal{N} \tag{2.1}$$

2.1.1 Filtered back projection

Using the Radon transform [124], we can represent an image $f(x, y)$ represented by the function as a series of line integrals at different offsets from the origin. This is shown in

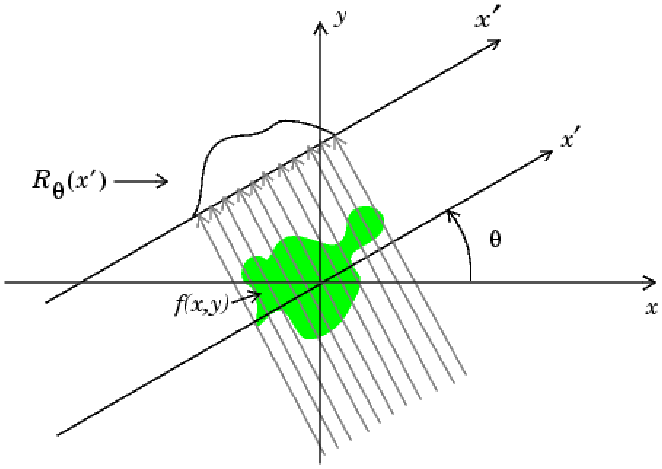


Figure 2.1: Radon Transform. Image is adapted from [41].

Figure. 2.1 and is defined mathematically as:

$$g(r,\theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\delta(x\cos\theta + y\sin\theta - r)dxdy \tag{2.2}$$

, where r is the perpendicular offset of the line of projection. In other medical imaging techniques such as Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET), data is acquired similarly by projecting a beam through an object. As a result of its characteristic sinusoidal shape after collecting Radon transform data at all angle θ, is referred tp as "sinogram" or (raw) projection data.

Once projection data $g(r,\theta)$ is obtained, we need to solve the inverse problem by finding $f(x,y)$. Fourier slice theorem states that 1D Fourier transform of the projection $g(r,\theta)$ is equal to the 2D Fourier transform of $f(x,y)$ evaluated at that angle θ or $G(0,\theta)$, parallel to that the slice.

$$G(\omega,\theta) = F(u,v)|_{u,v=\omega\cos\theta,\omega\sin\theta} \tag{2.3}$$

2D inverse transform of the image is

$$f(x,y) = \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(u,v)e^{j(ux+vy)}dxdy \tag{2.4}$$

After changing of variables,

$$f(x,y) = \frac{1}{4\pi^2} \int_0^{2\pi} \int_{-\infty}^{\infty} |\omega| G(\omega,\theta) e^{j(x\cos\theta + y\sin\theta)} d\omega d\theta \tag{2.5}$$

Here, we multiply the projection with $|\omega|$ in Fourier domain. This is called filtered back projection. Intuitively, $|\omega|$ is a ramp filter (high-frequency filter) to compensate low-frequency components accumulated during forward projection. This process corrects the image by reducing blurring. However, because the ramp filter emphasizes too many high-frequency components of the image, it can cause unwanted noise. Several other high-pass filters are commonly used to reduce noise. In medical physics, this is referred to as reconstruction kernels. The kernel affects the appearance of image structures by smoothing or sharpening the image. Different kernels have been developed for specific anatomical applications. There is generally a trade-off between spatial resolution and noise for each kernel. Smoother kernels produce images with less noise, but at the expense of reduced spatial resolution. A sharper kernel provides images with a higher spatial resolution, but increases the amount of noise in the image. It is important to select a reconstruction kernel based on specific clinical applications. When conducting exams for the brain or abdomen, for example, smooth kernels are typically used in order to reduce image noise and enhance low contrast detection. In contrast, sharper kernels are typically used to assess bone-like structures due to the clinical requirement for better spatial resolution.

The second dimension of reconstruction parameters is slice thickness, which controls the spatial resolution in the longitudinal direction. The medical professional's responsibility is to select the most suitable reconstruction kernel and slice thickness for each clinical application to minimize radiation dose and maintain image quality. Recommended protocols have been developed by the American Association of Physicists in Medicine (AAPM) Working Group to be used in the specific context of Lung Cancer Screening [2]. The protocols were based in part on manufacturer's guidelines, but were also adapted based on the National Lung Screening Trial.

2.1.2 Radiation dose

Dose, however, is not a reconstruction parameter but an acquisition parameter. Dose affects the noise level of images and thus the Signal-to-Noise Ratio (SNR). The radiation risk of X-ray CT gained increasing concern in the past decades. Studies have shown that although exposure to ionizing radiation from natural or background sources has not been changed since 1980, in the US the total per capita radiation exposure has nearly doubled, and experts believe the main reason is increased use of medical imaging [99]. Researchers at Brigham and Women's Hospital in Boston conducted a study in 2009 to estimate the risk of cancer caused by CT scans over the course of 22 years for 31462 patients. Overall, the increase in risk was 0.7% higher than the overall lifetime risk of cancer in the United States. On the other hand, patients with multiple CT scans were at higher risk, ranging from 2.7% to 12.7%. An increased risk of thyroid cancer and leukemia may also be associated with CT scans in adults and those diagnosed with non-Hodgkin lymphoma (NHL) at a young age. [109, 122].

Until we learn more, the general consensus is to limit the use of ionizing radiation for medical procedures. Therefore, the U.S. Preventive Services Task Force (USPSTF) recommends annual LDCT for lung cancer screening for people who have higher risks [52]. However, lowering CT scan dose leads to noisy raw data as well as streak artifacts after reconstruction. Because of the reduced X-ray dose, image noise may degrade image quality and result in an unsatisfactory diagnostic accuracy. Extensive studies have been conducted to minimize noise and artifacts for LDCT, including iterative reconstruction algorithms and image post-processing.

While the first CT scanners used iterative algorithms in 1970s, their clinical application was impeded by a lack of computational power [105, 112, 134]. Researchers have been developing new iterative algorithms for better LDCT image reconstruction [34]. They are followed by forward projection to the original reconstructed image to create simulated projection data, and then compared to the measured raw data. An updated reconstructed image is generated in case of a mismatch. The alternating process of correcting reconstructed imaging data and simulating projection data is repeated until a pre-defined condition is met and the final image is generated. The correcting process aims to optimize an objective function based on a system model, a statistical noise model, or prior information about the true image [27, 66, 100, 117]. Among the most popular image priors are total variation (TV) regularization [111], dictionary learning [137], and wavelet frames [20]. However, their computational cost and sensitivity to parameters changes restrict their practical application.

Methods of image post-processing after reconstruction are much more efficient than iterative reconstruction because they act directly on reconstructed images. The K-SVD method is proposed to reduce artifacts in CT reconstructions inspired by compressed sensing [7]. Another technique of practical use for CT imaging post-processing is Block Matching 3D (BM3D); it exploits similarities among the image blocks [22, 31, 54]. The Nonlocal Mean (NLM) filtering method estimates noise components based on multiple patches extracted from different locations in the image [70]. These post-processing methods have significantly improved the quality of the images; however, the results typically suffer from blurring and artifacts due to the nonuniform distribution of reconstruction noise.

Meanwhile, deep learning is becoming increasingly popular for computer vision tasks, but it is also proving to be very useful for denoising LDCT. The same concept is extended to computer vision tasks beyond denoising, such as super-resolution. Section 2.3 explores the related work that utilize deep learning for noise reduction and mitigates imaging variabilities due to multiple CT parameters.

2.2 Radiomics

Radiomics refers to a branch of medical science based on extracting a variety of pre-defined quantitative imaging features from radiographic images using data-driven algorithms. The features of these images tend to reveal disease characteristics that are not readily apparent to radiologists. It is believed that radiomics is capable of using the unique imaging characteristics of disease forms to predict prognosis and therapeutic response, thus facilitating personalized medicine [60]. Nevertheless, the technique can be applied to any medical study requiring medical imaging.

Radiomic features can be divided into shape-based features, first-order features, and higher-order texture-based features [93]. With the exception of shape features, all features can be calculated on either the original or derived images, obtained by applying preprocessing filters such as Wavelet, Lapacian of Gaussian (LoG) filters.

First-order features describe the distribution of voxel intensities within the region defined by the mask through first-order statistics. Still, they do not represent the voxel-level relationships within the image. The inter-voxel relationships can be captured by texture-based features, the gray level co-occurrence matrix (GLCM) , gray level run length matrix (GLRLM), gray level size zone matrix (GLSZM), and neighborhood gray zone difference matrix (NGTDM). GLCM features are the most commonly used textural features applied to various of medical imaging modalities to characterize biological tissue [6, 95, 116]. A gray level run length matrix (GLRLM) identifies the number of adjacent voxels with the same gray level value. It characterizes the gray level run lengths of different gray level intensities in any direction [113]. In addition, GLRLM-based features are also being used to describe biological tissues across other medical imaging modalities [55, 115]. A number of NGTDM features have been developed to correlate the quantitative values of texture features as closely as possible to human's visual interpretation of texture [21,123]. For example, coarseness provides a measure of local uniformity, while contrast provides quantitative information about differences between intensity levels in neighboring regions.

Using radiomic features, researchers suggested that a subset of intratumoral heterogeneity can improve survival prediction [6]. In addition, studies showed that radiomic features linked to prognostics in lung cancer might also be useful in head-and-neck cancer. However, Parmar et al. [95] demonstrated that some radiomic features might be associated with different prediction outcomes depending on the type of cancer. The researchers also observed that radiomic features are sensitive to the disease type. The features relevant to predicted lung cancer survival do not always successfully predict survival from head and neck cancers. Several studies have shown that radiomic features predict treatment response better than conventional measures, such as tumor volume and diameter [123]. Radiomics has also been shown to be beneficial in predicting a patient's immunotherapy response using pre-treatment PET/CT for Non-small-cell lung carcinoma (NSCLC) patients [91].

The availability of large CT datasets coupled with advances in medical image analysis has led to a proliferation of machine learning (ML) models that utilize quantitative image (radiomic) features for prediction and classification. However, few radiomic features are used in clinical practice. One significant barrier is that variations in CT acquisition and reconstruction impact downstream analyses, resulting in radiomic features with poor reproducibility. The inability to reproduce quantitative image features is well-documented [53,76,78,125,148]. Prior studies have demonstrated that differences in dose, slice thickness, and reconstruction kernel affects reproducibility in radiomics and tasks such as lung nodule detection and segmentation [28, 29]. Similarly, robust methods to mitigate the effects of CT scanner-specific acquisition and reconstruction parameters are critically needed to generate reliable radiomic features and achieve consistent algorithm performance. The following chapters will explore the impact of CT image normalization on radiomic feature value variability.

2.3 Related work

This section presents an overview of the related works for mitigating variations due to multiple CT parameters. CT voxel intensity profiles are created at different scales and shift factors due to CT parameter variations. Changes in the imaging intensity domain do not contribute equally to the quantitative imaging feature extraction, leading to biases. As a result, statistical analysis and machine learning based on radiomic features are notoriously sensitive to such changes, which subsequently hampers building robust models by pooling data. Thus, normalization is used prior to downstream image analysis. Given this context, there are two main workstreams to address this challenge: 1.) normalizing in image intensity domain directly and 2.) normalizing in quantitative imaging features domain (radiomics).

In the first workstream, acquisition protocols and reconstruction settings are normalized, for example, based on CT imaging guidelines. Although it can reduce variabilities due to multicenter effects, recent analysis has demonstrated that it cannot eliminate them [19]. However, recent developments based on GANs have shown great promise to generate images with a more similar appearance to the reference [47]. In the second workstream, to ensure that downstream statistical analysis only relies on robust features that are not affected by multicenter variations, only those features are selected. It is also possible to keep all derived features, but their statistical properties are reconciled (normalized) to have a common distribution. A variety of methods have been discussed in this category, such as ComBat or "combating batch effects when combining batches" [51].

The simplest form of normalization is standardization or centering. $f(x) = (x - \mu)/\sigma$, where x and $f(x)$ are the original and normalized feature values, respectively, μ and σ are the mean and standard deviation of the feature values. Feature values can either be extracted from quantitative imaging features or the original image intensity values. However, a simple standardization does not capture local intensity relationships that are essential for recovering image features for normalization.

2.3.1 Normalization in image intensity domain

Histogram matching-based algorithm has been widely used prior to ML-based techniques became widely available [131]. Histogram matching translates input images by mapping the source image's histogram to the target image's histogram using a prior cumulative distribution function (CDF). In practice, the target reference image's cumulative histogram is quite often missing or not well defined. The major disadvantage of histograms is the loss of information about local features within the image. However, it is also possible to divide a source image into patches and match histograms for every patch, hoping to achieve location-specific image synthesis using patch-based representations. Histogram matching based on patches, however, can introduce artifacts along the edges of patches. As Figure 2.2 shown, histogram matching is sensitive to parameters changes and artifact-prone.

Additionally, extracting abstract features beyond image intensity requires sophisticated preprocessing, e.g.. textural analysis. It is necessary to apply image processing filters prior to matching a histogram with an image. There are many ways to analyze images for texture analysis, such as applying Gabor filters to analyze specific frequency content in specific directions and a localized area. Feature engineering involves designing a filter to extract salient features from images. In summary, focusing only on the intensity information alone is insufficient for ensuring consistent image normalization.

The transforming relationship between unnormalized and target image is not necessarily linear. To ensure robust quantitative imaging features from translated images, models that consider nonlinear mappings are essential in order to mitigate the effects of variations from multiple dimensions. For example, several studies have been conducted to develop better image processing models for low-dose CT denoising thanks to the recent advancement of deep learning-based image translation techniques. The same approach can be extended to mitigation tasks beyond CT denoising. Those works fall into two categories: Convolutional neural network-based (CNN) and generative adversarial network-based (GAN) approaches,

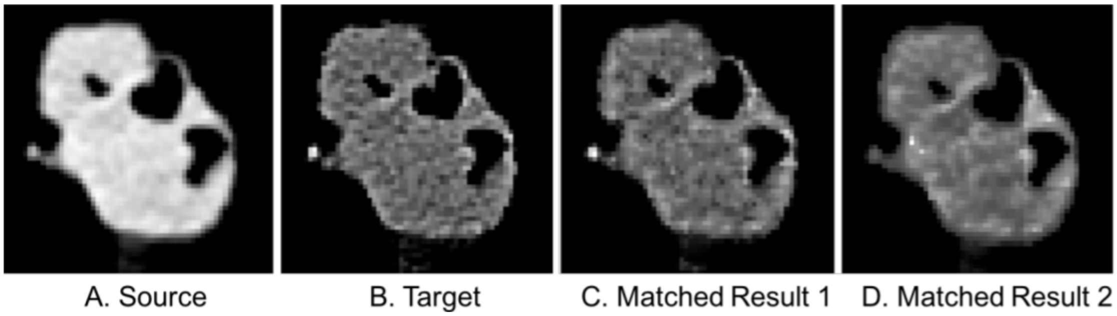


Figure 2.2: Histogram matching example. A.) unnormalized image. B.) reference image. C. and D.) normalized images using different histogram matching parameters (bins and patch size). Artifacts can be seen at the edges of C are blurred; D has a discontinuous region at the right of the tumor. Images are adapted from [72].

both seeking a function G_θ to map an input low-dose image x to a ground-truth routine-dose image y so that the density $\Pi(G_\theta(x))$ is close to the real data density.

2.3.1.1 CNN-based approaches

Mitigating the effects of dose has been the focus of multiple prior studies involving low-dose CT. CNN-based approaches utilize training on a deep convolutional neural network to optimize mean square error (MSE) loss, which directly maps a corrupted input image to a reference at a standard condition. The most commonly used ones are variants of U-net [102]. Chen et al. [17] used a residual encoder-decoder convolutional neural network (RED-CNN) to reduce noise in low-dose CT images. Figure 2.3 depicts the network structure of RED-CNN. RED-CNN is derived from the classic U-net [46] that contains an encoding and decoding branch with skipped connections.

Park et al. [94] used a U-net to overcome partial volume effects by learning an end-to-end mapping between 15 mm and 3 mm slice thickness images. Differences in reconstruction kernels can substantially impact texture features, which are commonly used in Computer Aided Detection (CAD) algorithms for disease characterization. Lee et al. [19, 65] used a

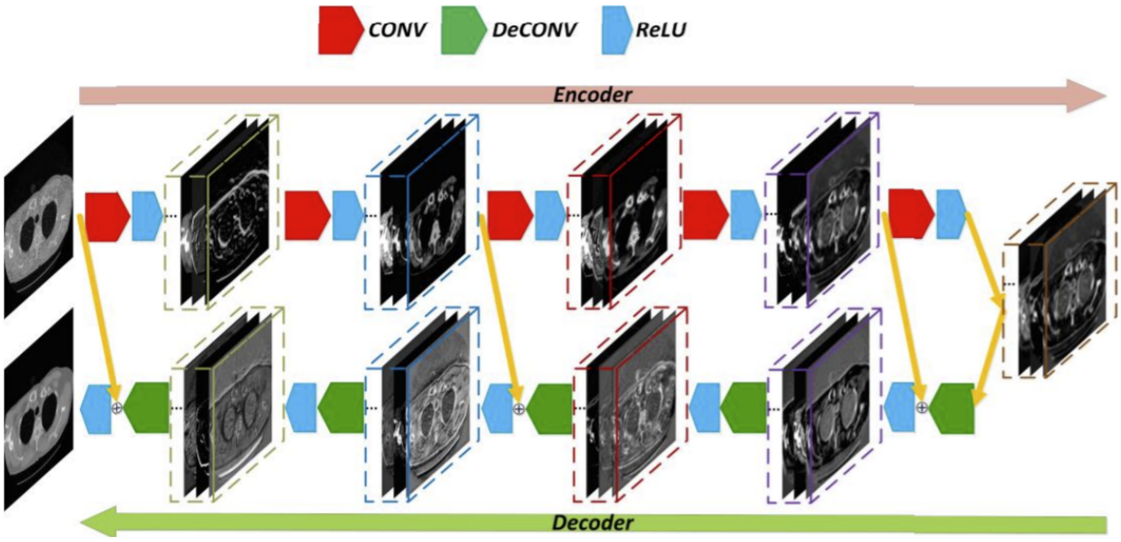


Figure 2.3: Architecture of RED-CNN network. Images are adapted from [17].

CNN to convert between smooth and sharp kernels. They demonstrated that normalizing kernels reduces variation in computed emphysema scores.

One issue with the CNNs is that they heavily rely on convolutional kernels, which integrate fixed-size filters to process one local neighborhood at a time. As a result, they do not efficiently retrieve information about a large region's structure. "Self-attention" was introduced in [67] to capture a wide range of spatial information both within CT slices and between CT slices. CNNs powered by the self-attention mechanism are able to leverage pixels that have more significant relationships regardless of their distance and achieve better denoising results.

Though these innovative network structures have produced impressive results, they learn the direct mapping between target and source, typically using the mean squared error (MSE) between the network output and the ground truth as the loss function. Despite its simplicity, MSE implies that the underlying image data was derived from a Gaussian prior, which is rarely the case for complex imaging data. Pixel-wise MSE results in over-smoothed edges and loss of detail. When an algorithm attempts to minimize per-pixel MSE, it ignores images with textural features essential for human visual perception. In the next section, we demonstrate the power of using GAN-based approaches to overcome the oversmoothing effect, as seen in CNN.

2.3.1.2 GAN-based approaches

Over the past few years, GAN-based approaches have been shown to generate realistic samples that are virtually indistinguishable from the training data. GANs take a minmax game theory approach that utilize adversarial training by directly sampling data and thus bypass any explicit density function estimation [36]. In the context of minmax games, there are two players, a generator G and a discriminator D. The basic formulation of GAN is presented in equation 2.6. D and G are trained by solving the following minimax problem.

$$\min_{G} \max_{D} \mathcal{L}(D, G) = \mathbb{E}_{x \sim P_r}[\log D(x)] + \mathbb{E}_{z \sim P_z}[\log(1 - D(G(z)))] \tag{2.6}$$

Here, $\mathbb{E}[\quad]$ denotes the expectation operator. P_r and P_z represent the real (reference) data distribution and the generated data distribution. z denotes a noisy feature vector that encodes a unique image. Generator G transforms z to an image sample that looks similar to the real one from a data distribution, denoted by P_g. If D were trained to behave like an optimal discriminator for G, the minmax optimization process is equivalent to minimizing the Jensen-Shannon (JS) divergence of P_r and P_g. The generative model framework can be extended to the "conditional" generative model, where one can impose a prior of unnormalized image x to the formulation (2.7), allowing G to transform z to a reference image sample $G(z; \hat{x})$ given a unnormalized image $\hat{x}$. An example of the GAN network structure is shown in Fig. 2.4.

$$\min_{G} \max_{D} \mathcal{L}(D, G) = \mathbb{E}_{x \sim P_r}[\log D(x)] + \mathbb{E}_{z \sim P_z, \hat{x} \sim P_{\hat{x}}}[\log(1 - D(G(z; \hat{x})))] \tag{2.7}$$

Wolterink [136] was the first to use a GAN to translate low-dose scans to appear similar to scans acquired at higher doses. Yang et al. [139] pursued a WGAN-based approach to enhance low-dose CT images, showing the effectiveness of a perceptual loss module to retain

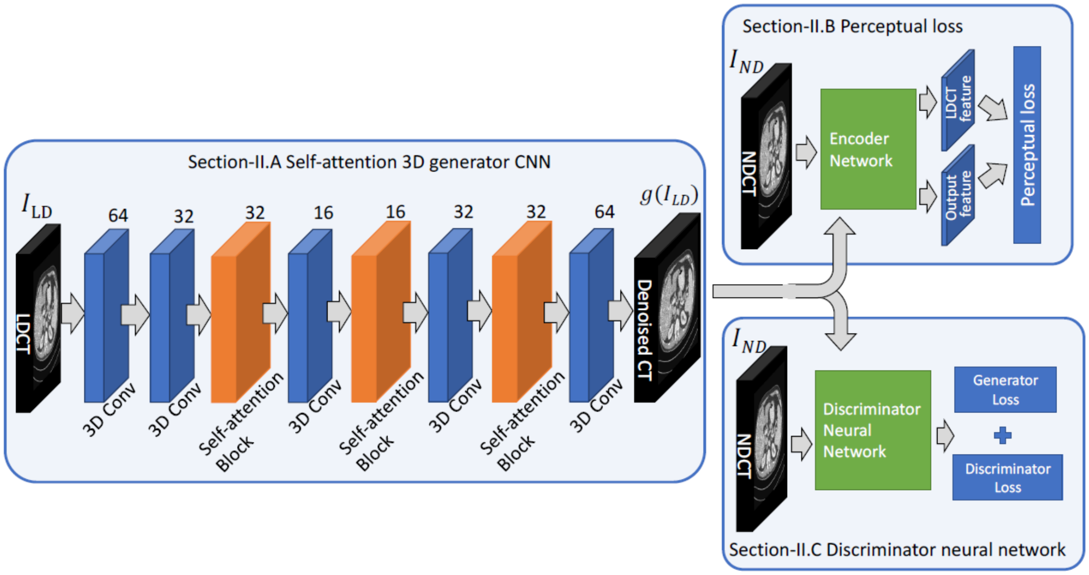


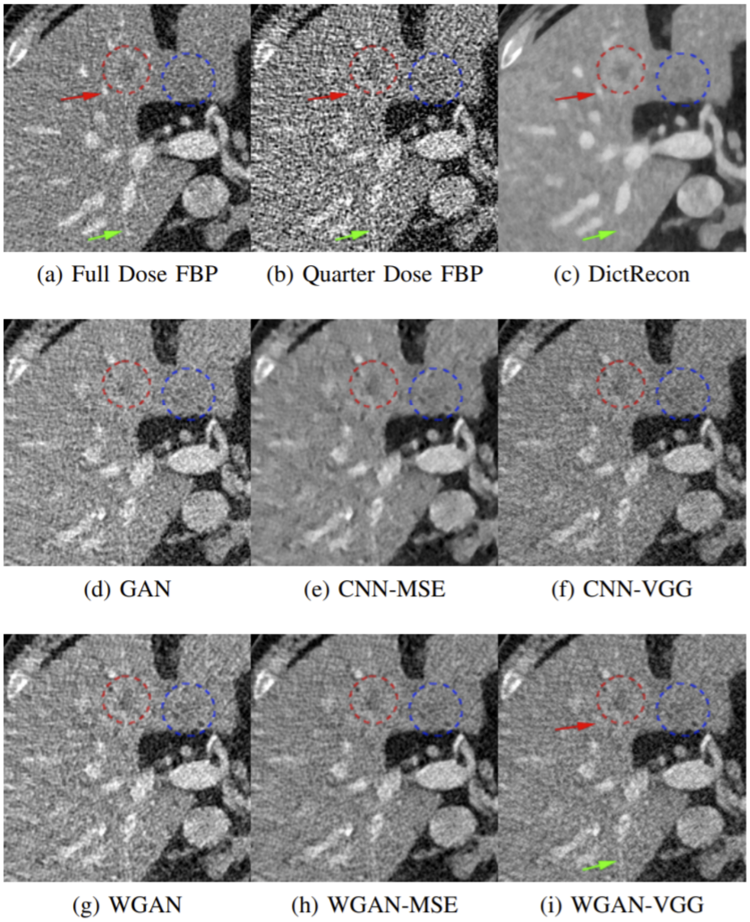
Figure 2.4: An example of Network architecture of GAN-based LDCT denoising model. Images are adapted from [67].

the texture characteristics of the image and avoid the oversmoothing effect. Figure 2.5 depicts a collection of normalized images for the AAPM dataset [3] study. It can be seen that GAN is able to recover textural details and thus improve image quality.

Wei [133] implemented a spectral-norm GAN (SNGAN) to stabilize GAN training and was able to mitigate variations due to multiple CT parameters. You et al. [144] demonstrated that a CycleGAN [151] could recover high-resolution images from down-sampled ones. Chen [18] implemented a similar GAN-based model for a MRI dataset. GAN-variants still need the use of content loss (such as MSE) to preserve structural consistency. Both approaches involve minimizing the MSE between the ground truth data.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \|G_\theta(x) - y\|^2 \tag{2.8}$$

Studies [49, 84, 152] have shown that GANs suffer from mode collapse and are prone to ignore the input noise vector z. All GAN-based methods mentioned above discourage using


(a) Full Dose FBP (b) Quarter Dose FBP (c) DictRecon

(d) GAN (e) CNN-MSE (f) CNN-VGG

(g) WGAN (h) WGAN-MSE (i) WGAN-VGG

Figure 2.5: DictRecon, CNN and GAN-based approaches evaluation results on AAPM LDCT challenge dataset. Images are adapted from [139].

the random vector z, and therefore the mapping is deterministic.

The limitations of existing CNN- and GAN-based works may be summarized as follows: a) they are typically trained on scans that have Gaussian noise added to them post-reconstruction, which does not reflect the physics of noise generation in raw projection at a low dose, b) most of the works only mitigate a single source of variability (e.g., dose or slice thickness) rather than address the impact of multiple parameters simultaneously, a scenario that is routinely encountered in practice, and c) they are evaluated using voxel-level metrics such as peak signal-to-noise ratio, which does not necessarily reflect how such normalization techniques would influence tasks such as emphysema scoring or nodule detection.

2.3.2 Normalization in QIFs domain

Combat is an empirical Bayesian method for data harmonization that was initially designed for genomic [85, 97, 110]. Various laboratories, tools and technicians may handle samples differently, leading to variations in measurement results. Johnson et al. [51] used the term "batch effect" to refer to the variabilities. Generally, ComBat applies to situations in which various features of the same type are measured for each subject, or where imaging-derived feature metrics are derived from different anatomical regions or voxels. "Batch effect" is conceptually related to variations in radiomic features due to differences in scanner models, acquisition protocols, and reconstruction settings across multiple centers [11, 81, 92]. As a result of ComBat, derived QIF data generated under different CT conditions can be represented in a common space, taking into consideration the effects of multicenter variations. ComBat works as described in formulation 2.9, where Y_{ijg} refers to extracted radiomics features values, and $X\hat{\beta}_g$ denotes covariate matrix for sample conditions and regression coefficients in the model that are not caused by CT parameter variability. Least-square is used to estimate features-wise mean and standard deviation, $\hat{\alpha}_g$ and $\hat{\sigma}_g$ for feature g, sample j, and group i.

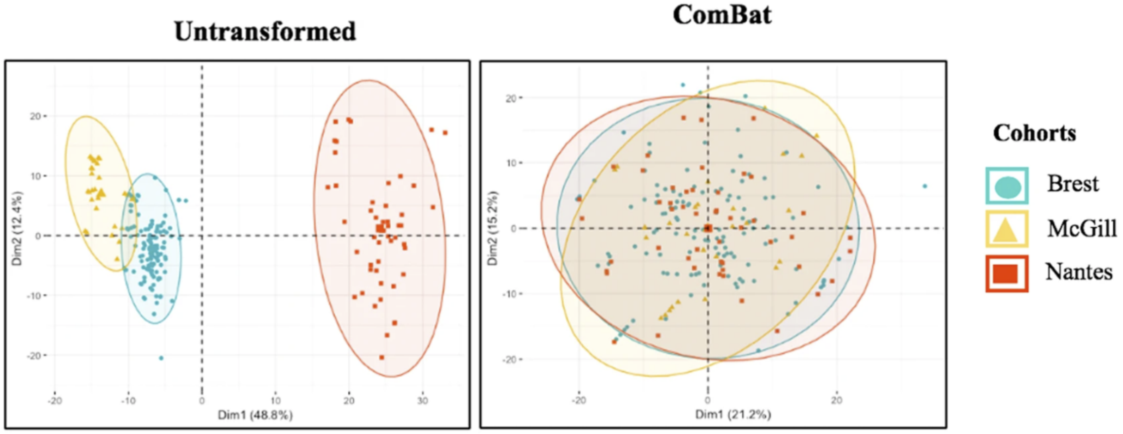


Figure 2.6: PCA and summary distribution in LACC dataset (Locally advanced cervical cancer): Scatter plots of two principal components of the radiomic features across the three groups (Brest, McGill, Nantes) using untransformed (unnormalized) data or data normalized with ComBat. Images are adapted from [103].

$$Z_{ijg} = \frac{Y_{ijg} - \hat{\alpha}_g - X\hat{\beta}_g}{\hat{\sigma}_g} \tag{2.9}$$

After standardization, Z_{ijg} is assumed to take the parametric forms for prior distribution, $\mathcal{N}(\gamma_{ig}, \delta_{ig}^2)$. Method of moments is used to estimate hyperparameters used to compute the empirical Bayes estimates of conditional posterior means features-wise for the center effects parameters [11, 51]. The final adjusted values after feature normalization are given by 2.10. As shown in Figure 2.6, ComBat is able to mitigate heterogeneities observed in radiomic features among three groups of data.

$$Y_{ijg}^{*} = \frac{\hat{\sigma}_g}{\hat{\delta}_{ig}^{*}}(Z_{ijg} - \hat{\gamma}_{ig}^{*})\hat{\alpha}_g + X\hat{\beta}_g \tag{2.10}$$

ComBat empirical Bayes estimation method has several advantages over a general linear model, including using a covariate to model a site or scanner as a fixed effect. Notably, ComBat is more resilient to outliners in cases of small sample sizes [51]. ComBat assumes that, for a given scanner, scanner effects across features originate from a common distribution, and leverages information across features to shrink estimates towards a common mean [11].

With the increasing popularity of deep learning models for medical analysis tasks, deep features are extracted instead of radiomic features. Generally, deep features cannot be interpreted, and are depend only on pre-trained models. Since deep learning models are not standard, each institute can develop models that vary from different disease types. Nevertheless, the model parameters change with each iteration of model training for a fixed model, resulting in different forms of deep features. Therefore, normalization in the QIF domain is no longer feasible. Although Combat is effective at normalizing radiomic features, it remains to be seen whether the method can also be applied to deep learning models.

CHAPTER 3

Conditional Generative Adversarial Networks

Variations in the appearance of Computed Tomography (CT) scans due to differences in acquisition and reconstruction parameters (e.g., dose, slice thickness, kernel) adversely impact the reproducibility of quantitative image features and the performance of machine learning models. Prior studies have attempted to mitigate the effect of a single parameter (e.g., dose), but variability in image appearance often results from the combined effect of differences in multiple parameters. In this study, we propose a 3D spectral-norm generative adversarial network (SNGAN) that normalizes CT images acquired and reconstructed under different conditions. We used raw projection data from low-dose chest CT scans of patients to generate a variety of reconstructions that simulate different dose levels, slice thicknesses, and reconstruction kernels. Each scan from 186 patients was reconstructed using 10 image conditions representing different acquisition and reconstruction parameters. Defining one condition to be the reference (i.e., a scan acquired at 100% dose, 1.0 mm slice thickness, medium kernel), we trained SNGAN models to normalize all other scans to the reference. We evaluated SNGAN against other state-of-the-art methods by comparing image quality metrics, impact on computed radiomic feature values, and a specific task (i.e., lung nodule detection). Our SNGAN improved perceptual similarity by 22%, compared to another GAN-based method. SNGAN resulted in smaller radiomic feature errors when compared to the reference condition (16% increase in features with “good” and “moderate” agreement based on concordance correlation coefficient). Performance of the existing nodule detection model was more consistent on scans normalized using SNGAN achieved compared to unnormalized scans (up to 75% improvement in concordance correlation coefficient). Collectively, these results demonstrate the SNGAN's ability to normalize heterogeneous CT images and reduce adverse impacts on downstream tasks.

3.1 Introduction

We present a generative adversarial network (GAN)-based approach to mitigate the effects of working with CT scans that have been acquired and reconstructed using a range of parameters. Our goal is to attain consistent radiomic feature values and computer-aided diagnosis (CAD) performance when characterizing the same imaging abnormality across a wide range of input scans. We demonstrate this application in the context of lung nodule detection on low-dose CT. We hypothesize that the improved image quality results in better consistency in nodule detection. Our approach takes CT scans, acquired at varying doses and reconstructed using different slice thicknesses and kernels, as input and generates images that appear as if they were acquired using an identical set of parameters (i.e., a reference condition).

The novel CT image normalization method is based on a 3D GAN that utilizes a spectral-normalization technique. Our model, called spectral-normalization GAN or SNGAN, provides an end-to-end solution for normalizing scans acquired using different doses, slice thicknesses, and reconstruction kernels. Our work is distinctly different to "one-size-fits-all" data augmentation approach that makes robust and generalized models for different input conditions. Instead, we created separate models for each common CT condition that are used in clinical practice to map the unnormalized CTs to a reference. Separate models are more flexible to address certain input conditions where mitigation is needed. Each institution has its own set of commonly used protocols for acquisition and reconstruction. As prior work from our team suggests, certain parameter variability may not have an impact on the reproducibility of some radiomic features [30]. Thus, normalization might not be necessary for those conditions. However, one can choose the appropriate model for a particular condition when normalization is beneficial. The contributions of our work are as follows:

1. We trained and validated our SNGAN model on a large collection of raw projection data acquired from patients undergoing lung cancer screening using a high-throughput reconstruction and analysis pipeline [44].

2. SNGAN mitigates variability from three CT parameters simultaneously (dose, slice thickness, and kernel). 3D convolutions are used to improve the spatial resolution along the z direction.

3. A multi-pronged evaluation is conducted to evaluate the impact of SNGAN in a) optimizing per-voxel and perceptual metrics, b) reducing variability of computed radiomic feature values across different scans, and c) achieving more consistent performance of a previously published lung nodule detection [14] when executed on scans acquired using different parameters.

While using GANs to mitigate the effect of multiple parameters has not been widely explored, we compare our approach to one existing work called GANai [72]. GANai uses a conditional GAN (cGAN) and alternative training strategies to map two reconstruction parameters (kernel and slice thickness) to a target condition. However, its ability to improve the resolution along the z direction was fundamentally limited due to the 2D network structure. Their training was performed in 2D with source (thick slice) and target (thin slice) image pairs. The partial volume effect was not fully addressed. We overcome this by utilizing 3D convolutions in our network. Our method learns the spatial correlation in adjacent slices from thin slice scans and results in high z resolution in normalized images. The authors also did not investigate the impact of their cGAN on radiomic feature values due to variations in dose. To the best of our knowledge, no existing work has attempted to simultaneously normalize multiple CT parameters (dose, kernels and slice thickness). Moreover, while most studies evaluate their approaches using per-voxel image quality metrics or human reader studies, we designed the experiments to be objective and task-oriented, providing a clearer understanding of what impact the method would potentially have in practice.

3.2 Method

GANs have been used to generate photo-realistic images, translate images from one style into another, and perform image enhancement such as denoising [36, 64]. However, GANs are notoriously difficult to train and sensitive to the choice of hyperparameters. Arjovsky et al. [9] was the first to realize that a GAN's loss function had fluctuating gradients, frequently resulting in unstable training. Based on this finding, multiple studies attempted to identify better loss functions that achieved smoother landscapes. One potential function is the Earth-Mover distance used in the Wasserstein GAN with Gradient Penalty (WGAN-GP) [38]. WGAN requires the computation of the second-order gradient, thus increasing complexity. Our model utilizes a robust spectral-norm layer to satisfy the Lipschitz constraint as opposed to using Wasserstein distance.

As a comparison, we implemented a baseline 3D model based on a convolutional neural network (CNN) with absolute error ($\mathcal{L}_1$) loss. When specifically mitigating the effect of dose, we implemented WGAN-GP as a comparison. First, we evaluated image quality using the peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) [147] calculated along the axial, coronal, and sagittal planes. Second, we studied the effect of normalization on radiomic features by computing these features from segmented nodule regions using CNN and SNGAN. We calculated the absolute error and concordance correlation coefficient (CCC) between radiomic features computed using the scans acquired at the reference condition and unnormalized vs. normalized scans. Third, we fed the normalized images into an existing nodule detection algorithm, comparing the performance of our nodule detection algorithm when given unnormalized and normalized scans.

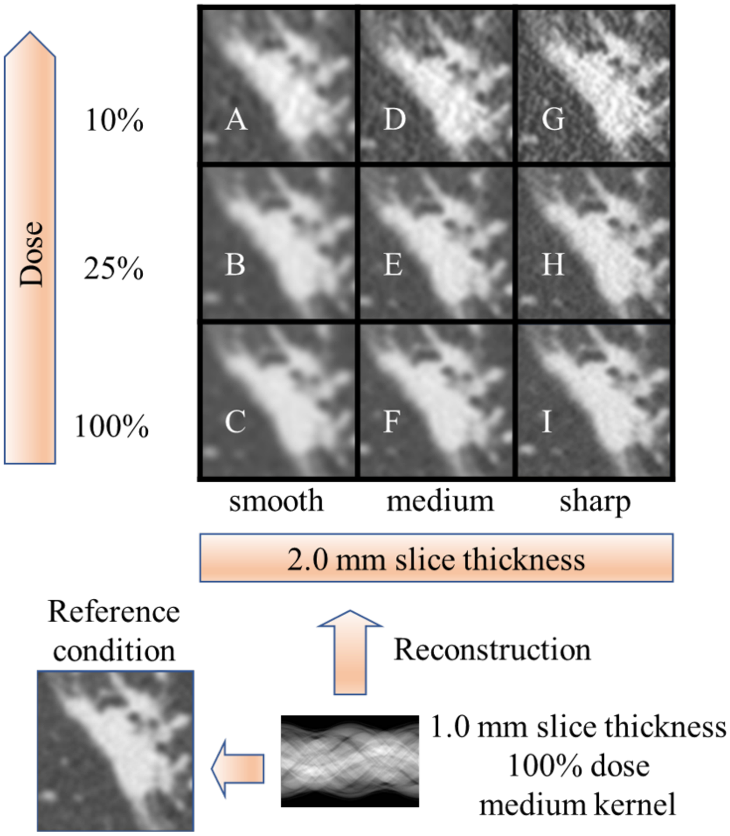


Figure 3.1: UCLA dataset. Condition A-I were generated from the reference raw sinogram data using dose simulation, followed by image reconstruction with weighted filtered back projection (wFBP) algorithm. [45]

3.2.1 Datasets

Our dataset consisted of 186 chest LDCT exams that were acquired at an equivalent dose of ~2mGy. The images were collected under IRB/ethics board protocol number 11-000126 (Computer Analysis of CT images). Raw projection data of scans performed on Siemens CT scanners (Definition Flash, Sensation 64, Definition AS) were exported. Poisson noise was introduced into the raw projection data, as described in [44] at levels that were equivalent to 10% and 25% of the original dose. Original full-dose and reduced dose projection data were then reconstructed into an image size of 512 × 512 using three reconstruction kernels (smooth, medium, sharp) and two slice thicknesses (1.0 mm, 2.0 mm). Fig. 3.1 illustrates how a cropped image of a lung nodule would appear across each of the reconstructed conditions. The dataset was split as follows: 80 scans for training, 20 scans for validation, and 86 scans for testing. In the test set, 43 scans (50%) contained a total of 68 lung nodules. The centroids of these nodules were marked by a trained image analyst using the original radiologist report as a reference. In this study, the reference condition was set to be 100% dose, medium kernel, and 1.0 mm slice thickness. The slice thickness and kernel was chosen to reflect parameters that are currently recommended for lung cancer screening [1].

3.2.2 Network architecture

3.2.2.1 GAN and loss functions

GANs consist of a generator G and a discriminator D. The generator learns a function $G(x)$ that takes an inputted scan x, outputting an image $\hat{y}$ that mimics the appearance of a scan y acquired under a reference condition. A discriminator D is trained using both the normalized scans generated by generator and actual scans acquired at the reference condition as inputs to differentiate between $\hat{y}$ and y. D constantly judges the similarity between $\hat{y}$ and y to improve the performance of the generator. A good generative model is achieved when the discriminator can no longer distinguish between them. In this work, we follow a similar network architecture as reported by Wei [133]. Inspired by Enhanced Deep Residual Networks (EDSR) [73], we use a SRResNet as our generator, which consists of multiple layers of residual blocks to extract features. The basic residual block is composed of two $3 \times 3 \times 3$ convolutional layers and a ReLU. The deep features extraction is followed by an upsampling block in the z-direction and two sequential convolutional blocks before outputting a generated normalized image. This upsampling block mitigates the partial volume effect resulting from thicker slices. Hinge loss is used for the discriminator to constrain D to focus on samples that are difficult to classify as model outputs versus actual scans. The generator loss function contains an $\mathcal{L}_1$ content loss and an adversarial loss. The discriminator loss $V_D(G, D)$ and generator loss $V_G(G, D)$ are shown in equations 3.1 and 3.2, where p_x and p_y are distributions of non-reference input scans and the reference scan, respectively. Training proceeds with alternating D and G updates, $\min_G \max_D[V_D(G, D) + V_G(G, D)]$, where Θ and W are the network parameters of the discriminator and generator, respectively. The weights of Θ and W are initialized using Kaiming initialization [42] with a scale of 0.1. The network structures are illustrated in Fig. 3.2. The baseline CNN model has the same architecture as the generator network but trained without the discriminator model. Either $\mathcal{L}_1$ or $\mathcal{L}_2$ loss can be used to train a CNN model. Details regarding the choice of loss functions can be found in 3.3.1.1.

$$\begin{aligned} V_D(G, D) = &\mathop{\mathbb{E}}_{y \sim p_y}[\min(0, -1 + D_\Theta(y)] \\ &+ \mathop{\mathbb{E}}_{x \sim q_x}[\min(0, -1 - D_\Theta(G_W(x)))] \end{aligned} \tag{3.1}$$

$$\begin{aligned} V_G(G, D) = &-\alpha_1 \mathop{\mathbb{E}}_{x \sim q_x}[D_\Theta(G_W(x))] \\ &+ \alpha_2 \mathop{\mathbb{E}}_{\substack{x \sim q_x \\ y \sim p_y}} \|G_W(x) - y\|_1 \end{aligned} \tag{3.2}$$

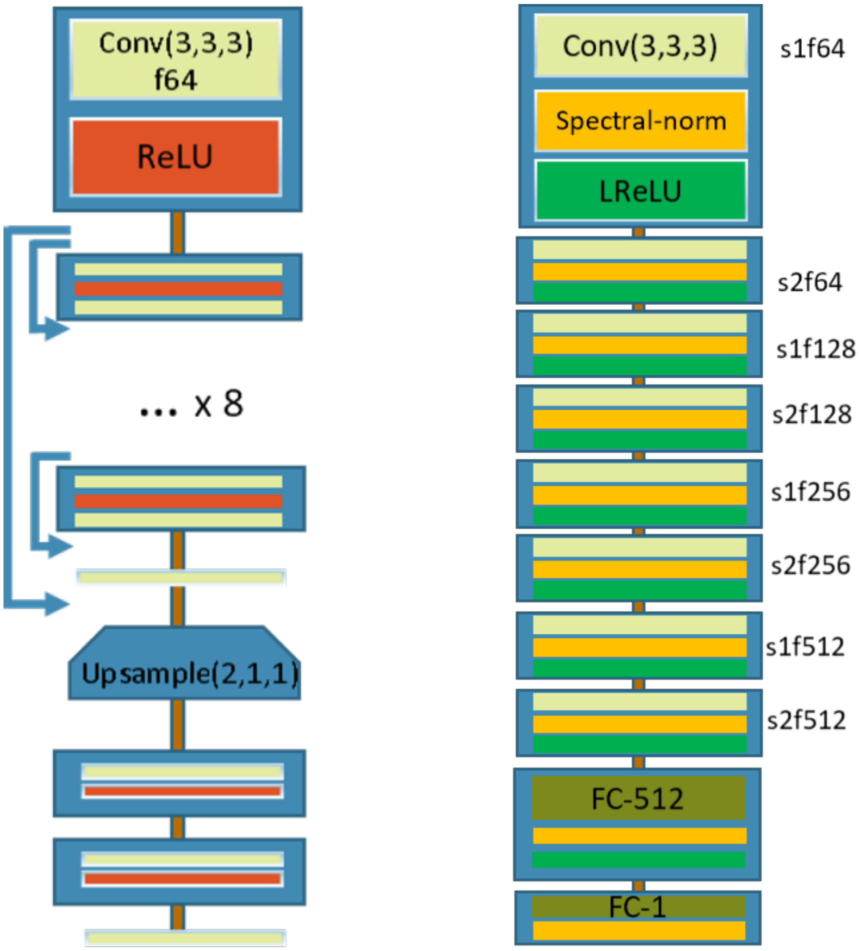


Figure 3.2: Network architecture of the generator (left) and discriminator (right). s{}f{} stands for stride number and filter number. The upsampling block only operates in the z direction.

3.2.2.2 Spectral-norm

To improve the training stability, WGAN-GP [38] imposed local regularization on the discriminator to satisfy the Lipschitz continuity constraint by penalizing the gradients. Here, we used spectral-norm [90] to achieve the same goal. The spectral-norm is a robust global regularization, as opposed to calculating the computationally expensive second-order gradient penalty term in WGAN-GP. Given a weight matrix $\mathbb{W}$, $\mathbf{K}$-Lipschitz constraint states that $\|\mathbb{W}x\| \leq \mathbf{K}\|x\|$, for any x and finite $\mathbf{K}$. The spectral-norm is the maximum singular value of matrix $\mathbb{W}$. One can normalize $\mathbb{W}$ by the largest singular value of $\mathbb{W}^T\mathbb{W}$ or $\sqrt{\lambda_1}$ to satisfy $\mathbb{1}$-Lipschitz continuity, where λ_1 is the dominant eigenvalue. However, computing eigenvalues using singular value decomposition (SVD) is not desirable for larger matrices due to the computational cost. We followed the same strategy introduced in [90] using power iteration to compute this value efficiently. The algorithm is outlined in Algorithm 1. Instead of performing multiple iterations to find the converging spectral-norm value, we only conducted a single iteration at each backward propagation, which reduced the computational complexity. Therefore, normalizing $\mathbb{W}$ by spectral-norm was seamlessly integrated into the global network parameter update step.

Algorithm 1: Spectral-norm power iteration algorithm.

```
Sample a random vector $\tilde{u_n} \in \mathscr{R}^{m_n}$ from an isotropic distribution, where $m_n$ is the
  dimension of the n-th layer;
for each mini batch do
    for each linear operation layer n do
        • $\tilde{v}_n \leftarrow \frac{\mathbb{W}_n^T \tilde{u}_n}{\|\mathbb{W}_n^T \tilde{u}_n\|}$;
        • $\tilde{u}_n \leftarrow \frac{\mathbb{W}_n \tilde{v}_n}{\|\mathbb{W}_n \tilde{v}_n\|}$;
        • Spectral-norm for layer n weight matrix,
          $\mathbb{W}_n^* \leftarrow \mathbb{W}_n / \sqrt{\lambda_1}$ where $\sqrt{\lambda_1} = \tilde{u}_n^T \mathbb{W}_n \tilde{v}_n$;
        • Update $\mathbb{W}_n$, with learning rate $\alpha$, and $\beta_{1,2}$
          using Adam optimizer,
          $\mathbb{W}_n \leftarrow \mathbb{W}_n - Adam(\nabla_{\mathbb{W}_n} \mathscr{L}(\mathbb{W}_n^*, D), \alpha, \beta_{1,2})$;
    end
end
```

3.2.3 Model training

We inputted patches of size 16×64×64 voxels (depth, height, width), excluding patches that were primarily outside of the body. We randomly generated the input patches to avoid overfitting. The raw voxel Hounsfield values were cropped to [-1000, 500] and were scaled to [0, 1]. The outputted patch from the network had a dimension of 32×64×64. As reported in [90], the best choice of the D/G update ratio was 1, which was what we used for our model. An Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ was employed. For the generator loss function, $\alpha_1 = 1$ and $\alpha_2 = 5e-3$ were used. These values were determined based on a grid search performed on the validation set. The batch size was set to 16. Training was stopped at 100k iterations when good convergence of image quality metric was achieved. The learning rate was $1e-5$ and halved every 20k iterations. We used a single Nvidia Tesla v100 GPU to train the model, taking approximately 60 hours. The `NVIDIA APEX` [88] mixed-precision training package was used to accelerate training and reduce GPU memory requirements. For robust inference and saving GPU memory, volumetric scans were represented using half precision (FP16) and split into smaller 3D patches of size 512×512×32 (height, width, depth). The outputted 3D patches are put back together with an overlap of 4 voxels in the z-direction to reduce artifacts in stitching.

3.2.4 Experimental design

We investigated two normalization scenarios: **1**) normalizing scans that were acquired at different doses but with the same slice thickness and kernel (i.e., single CT parameter normalization) and **2**) normalizing scans acquired at different doses, slice thicknesses, and reconstruction kernels (i.e., multiple CT parameter normalization). In scenario **1**, the network acted as a denoising algorithm, a special case of scenario **2**, in order to compare our SNGAN approach compared to another GAN-based denoising algorithm, WGAN-GP. The approach for each scenario is described briefly:

1. We trained models to normalize an image acquired at a simulated 10% dose to an image acquired at 100% dose. Image quality was evaluated using metrics such as PSNR, SSIM and LPIPS. As a comparison, we re-implemented WGAN-GP [139] with the following modifications: a) In the generator, we removed the ReLU in the last layer. We refer to this generator network as vanilla CNN. b) In the discriminator, the perceptual loss module was omitted given that a pretrained 3D perceptual network was not readily available.

2. Models were trained to simultaneously normalize differences in dose, kernel, and slice thickness. We trained individual SNGAN and baseline CNN models to perform a mapping between nine different image conditions (2.0 mm slice thickness scans acquired at 10% dose, 25% dose, and 100% dose, each reconstructed with either a smooth, medium, or sharp kernel) to the reference condition.

3.2.4.1 Image quality assessment

Our primary focus is to improve consistency in downstream analysis tasks, measured by CAD algorithm performance. Nevertheless, we performed evaluations using standard image metrics to provide a comparison to prior methods. PSNR and SSIM (shown in formula 3.3 and 3.4) are commonly used to measure local differences between the output (e.g., normalized image) and a reference image. However, these per-voxel metrics are computed using low-level features that may not reflect the types of higher-level features that inform specific tasks. Optimizing the loss corresponding to these metrics (e.g., using mean-squared error) leads to overly smoothed images and eliminates texture details [150]. To better assess the image quality, we used LPIPS, a perceptual metric that utilizes a pretrained VGG network to generate similarity scores from high-level feature space between two images. A lower LPIPS value represents a closer distance to the reference image. For each metric, results

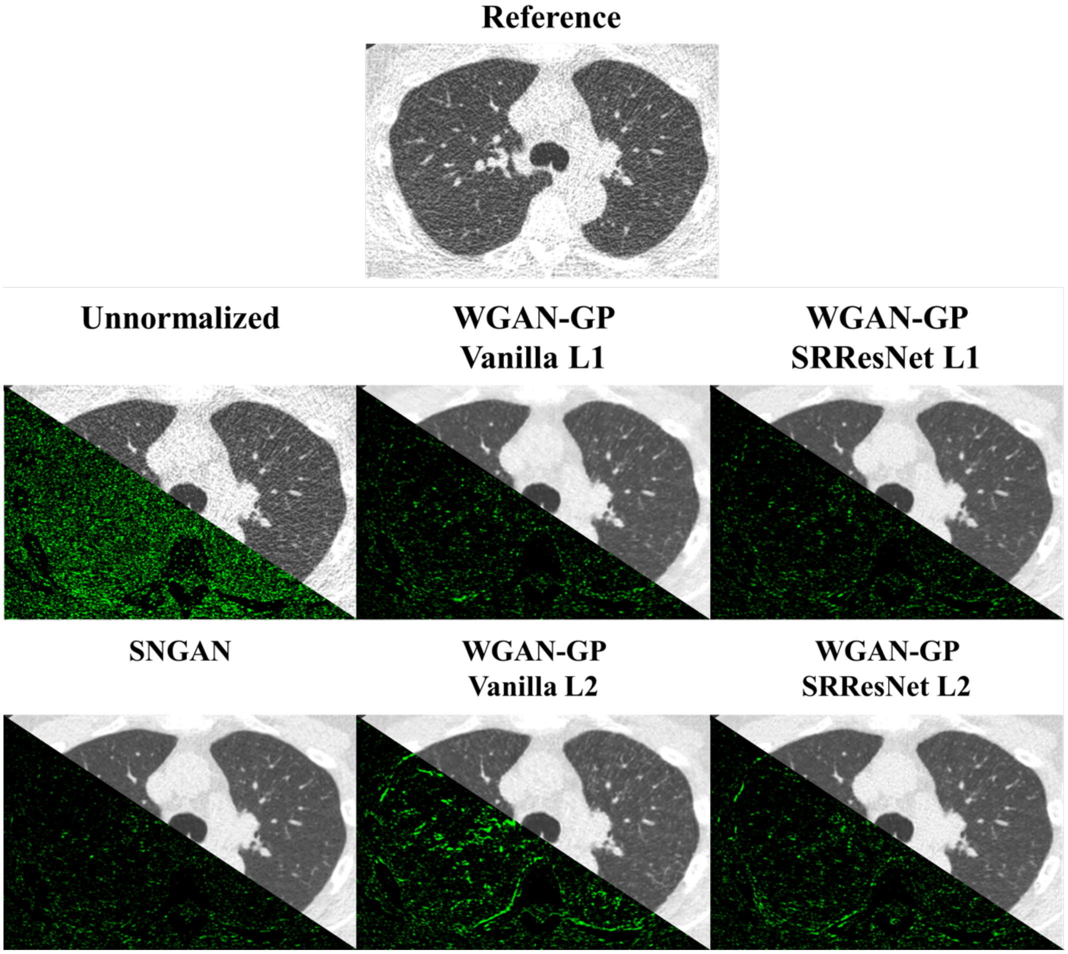


Figure 3.3: Single CT parameter (dose) normalization results. Difference images (bottom left) generated from the normalized showing the residual between unnormalized, WGAN-GP, and SNGAN compared with the reference condition.

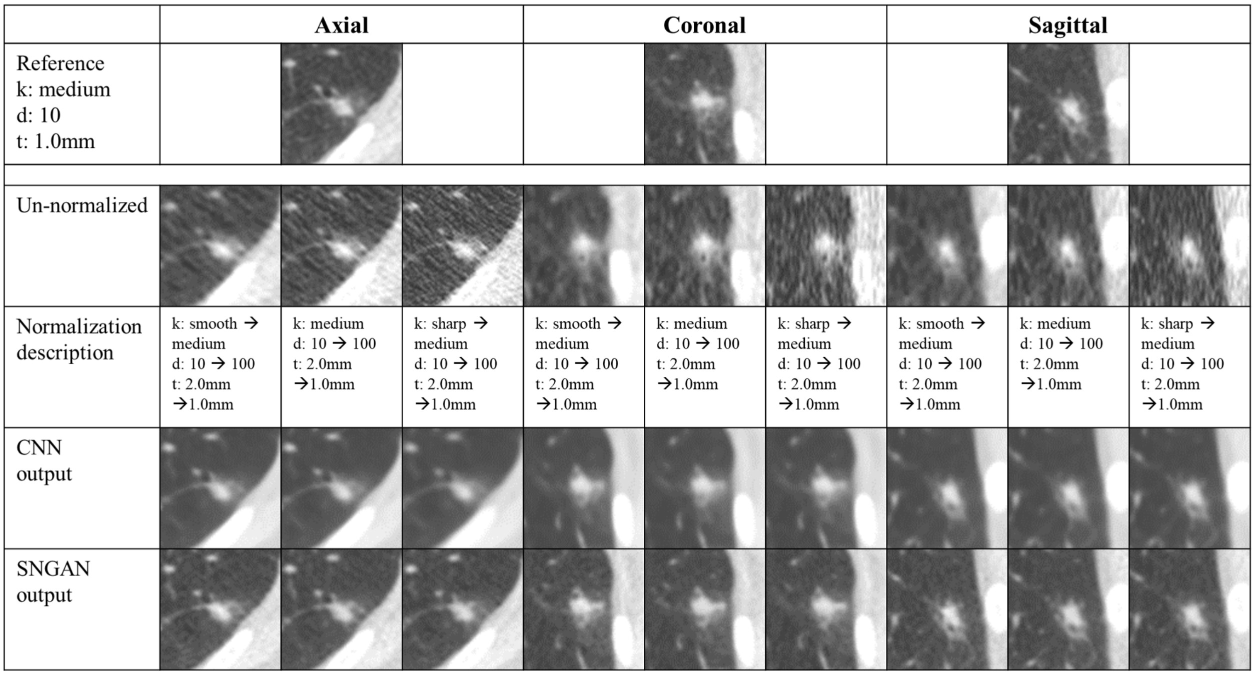

Axial
Coronal
Sagittal
Reference
k: medium
d: 10
t: 1.0mm
Un-normalized
Normalization description
k: smooth → medium d: 10 → 100 t: 2.0mm →1.0mm
k: medium d: 10 → 100 t: 2.0mm →1.0mm
k: sharp → medium d: 10 → 100 t: 2.0mm →1.0mm
k: smooth → medium d: 10 → 100 t: 2.0mm →1.0mm
k: medium d: 10 → 100 t: 2.0mm →1.0mm
k: sharp → medium d: 10 → 100 t: 2.0mm →1.0mm
k: smooth → medium d: 10 → 100 t: 2.0mm →1.0mm
k: medium d: 10 → 100 t: 2.0mm →1.0mm
k: sharp → medium d: 10 → 100 t: 2.0mm →1.0mm
CNN output
SNGAN output

Figure 3.4: Multi-parameter normalization results. A volume of interest (VOI) containing a lung nodule is displayed in axial, coronal, and sagittal plane. For a selected viewing plane, each column represents a normalization condition.

were calculated along the axial (x-y), coronal (x-z), and sagittal (y-z) planes.

$$PSNR(x, y) = 10 \log_{10} \frac{Max(\hat{y}_{ij}, y_{ij})^2}{\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(\hat{y}_{ij} - y_{ij})^2} \tag{3.3}$$

$$\begin{aligned} SSIM(\hat{y}, y) &= \frac{(2\mu_{\hat{y}}\mu_y + C_1) + (2\sigma_{\hat{y}y} + C_2)}{(\mu_{\hat{y}}^2 + \mu_y^2 + C_1)(\sigma_{\hat{y}}^2 + \sigma_y^2 + C_2)} \\ C1, C2 &\quad \text{constants to stabilize the division} \\ \mu_{\hat{y}}, \mu_y &\quad \text{mean of } \hat{y}, y \\ \sigma_{\hat{y}}, \sigma_y, \sigma_{\hat{y}y} &\quad \text{standard deviation and co-variance of } \hat{y}, y \end{aligned} \tag{3.4}$$

3.2.4.2 Radiomic feature analysis

To evaluate the impact of our SNGAN normalization approach on radiomic feature values, we used `pyradiomics` [127] to extract first-order, gray level concurrence matrix (glcm), gray level run length matrix (glrlm), and gray-level size zone matrix (glszm) features. Given that features were generated from 24×48×48 volumes of interest (VOI) containing the nodule, shape features were not computed. All features aforementioned can be calculated either on the original normalized images or preprocessed by applying: a) Laplacian of Gaussian (LoG) filters (σ=1,2,3,4,5 mm) and b) Wavelet filters {H, L} applied in the x, y and z directions (H: High pass, L: Low pass). We computed the absolute feature errors under each condition. The absolute error for the i-th feature was defined as $E_i = |X'_i - X_i|$, where X is a feature vector extracted from a VOI at some condition and X' is the feature vector from a VOI at the reference condition.

To facilitate the analysis of the high-dimensional features, we used t-Distributed Stochastic Neighbor Embedding (t-SNE) [80] to reduce feature vectors $[X_1^T, X_2^T, ...X_n^T]$ into two dimensions. For each of the nine mapping conditions (Scenario 2), we visualized the distribution of features extracted from unnormalized, reference, and normalized (CNN and SNGAN processed) scans. Kernel density estimation was used to estimate the probability distribution for each group from the corresponding data points.

We also used Lin's concordance correlation coefficients (CCC) [62] to compute the level of agreement between outputs generated from unnormalized and normalized scans for each mapping condition. The computed CCC matrix was visualized using a heatmap. In this study, a threshold of 0.9 was defined as having "good" agreement, a threshold between 0.9 and 0.8 was considered "moderate" agreement, and a threshold of 0.8 and below was considered "poor" agreement based on prior interpretations of the CCC [63, 138].

3.2.4.3 Lung nodule detection task

We evaluated the effect of using CNN- and SNGAN-based techniques to normalize scans prior to inputting them into a lung nodule detection algorithm. For this task, we utilize an existing algorithm [14], which utilizes common image analysis techniques such as intensity thresholding, Euclidean distance, and watershed segmentation to identify nodules in the test set. The detection model was trained using data from the Lung Database Image Consortium [10]. The focus of this analysis is not on the detection algorithm's absolute performance but on the relative impact of normalization. As such, instead of reporting absolute sensitivity and F1-score for the algorithm, we used CCC to measure the agreement of the model's sensitivity and F1-score between unnormalized or normalized (SNGAN, CNN) scans and the reference scans. Subject-level sensitivity and F1-score for scan i were computed by formula 3.5, 3.6, and 3.7 after running the algorithm on all test subjects across all image conditions. The CCC for subject-level sensitivity and F1-scores were then computed by 3.8 to evaluate the relative level of agreement in detection performance.

$$\text{sensitivity}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{3.5}$$

$$\text{precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \tag{3.6}$$

$$\text{F1}_i = \frac{2 \times \text{precision}_i \times \text{sensitivity}_i}{\text{precision}_i + \text{sensitivity}_i} \tag{3.7}$$

$$\text{CCC} = \frac{2\rho\sigma_{\hat{y}}\sigma_y}{\sigma_{\hat{y}}^2 + \sigma_y^2 + (\mu_{\hat{y}} - \mu_y)^2},$$
$$\text{where } \sigma, \mu = mean(\{\text{F1}_i\}), std(\{\text{F1}_i\}), \tag{3.8}$$
$$\text{and } \rho \text{ is the correlation coefficient.}$$

3.3 Results

To illustrate the results of Scenario 1, a single axial slice is shown in Fig. 3.3 comparing the unnormalized scan with different normalization techniques. Fig. 3.4 depicts the results of Scenario 2 for the same axial, coronal, and sagittal slices. All displayed lung window is centered at -600 Hounsfield units (HU) with a width of 1500 HU.

3.3.1 Scenario 1: Single CT parameter normalization

3.3.1.1 Image quality assessment

Table 3.1 summarizes the image quality results for mitigating the effects of different doses. SNGAN achieves superior PSNR, SSIM, and LPIPS compared to WGAN-GP. On average, compared to the of WGAN-GP with c loss, SNGAN improved metrics by 9.0%, 7.3%, and 22.2% respectively. To further understand what aspect of SNGAN was driving the improvements in metrics, we replaced the vanilla CNN generator in WGAN-GP with the SRResNet generator. We also trained the models with both $\mathcal{L}_1$ and $\mathcal{L}_2$ loss to see the impact of the choice of the content loss function. Fig. 3.3 shows each individual model's residual when comparing to the reference condition. SNGAN resulted in images that were less noisy as evidenced by the smaller number of residual pixels. For WGAN-GP approaches (b,c,e,f), the residuals were substantial along the chest wall, indicating that WGAN-GP struggled to minimize the error in the region where high-frequency components dominated. The convergence on Wasserstein distance for the various generator and loss functions are shown in Fig. 3.5. Our ablation analysis showed that a) due to multiple residual connections, the

SRResNet generator achieved superior Wasserstein distance convergence and was capable of generating images with better visual quality; b) the use of $\mathcal{L}_1$ loss improved the overall image quality and convergence; and c) even when WGAN-GP was equipped with the same generator architecture and $\mathcal{L}_1$ loss function, SNGAN still achieved superior performance across all three image metrics, indicating that SNGAN's discriminator played an important role in enhancing image quality.

Table 3.1: Image quality comparison. Results at **Ax**(ial), **Co**(ronal), and **Sa**(gittal) viewing planes were are calculated below. ↑ The higher the better: PSNR and SSIM. ↓ The lower the better: LPIPS.

GAN type		SNGAN	WGAN-GP			
Generator		SRResNet	Vanilla		SRResNet	
Loss function		$\mathcal{L}_1$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_1$	$\mathcal{L}_2$
PSNR	Ax	**31.09**	30.09	28.52	30.46	29.96
	Co	**32.76**	31.67	29.85	32.10	31.57
	Sa	**31.37**	30.38	28.99	30.73	30.28
SSIM	Ax	**0.7967**	0.7768	0.7456	0.7871	0.7741
	Co	**0.7771**	0.7500	0.7199	0.7675	0.7534
	Sa	**0.8002**	0.7758	0.7467	0.7915	0.7788
LPIPS	Ax	**0.1556**	0.1873	0.1919	0.16384	0.1691
	Co	**0.1548**	0.1982	0.2076	0.1655	0.1720
	Sa	**0.1362**	0.1744	0.1813	0.1454	0.1513

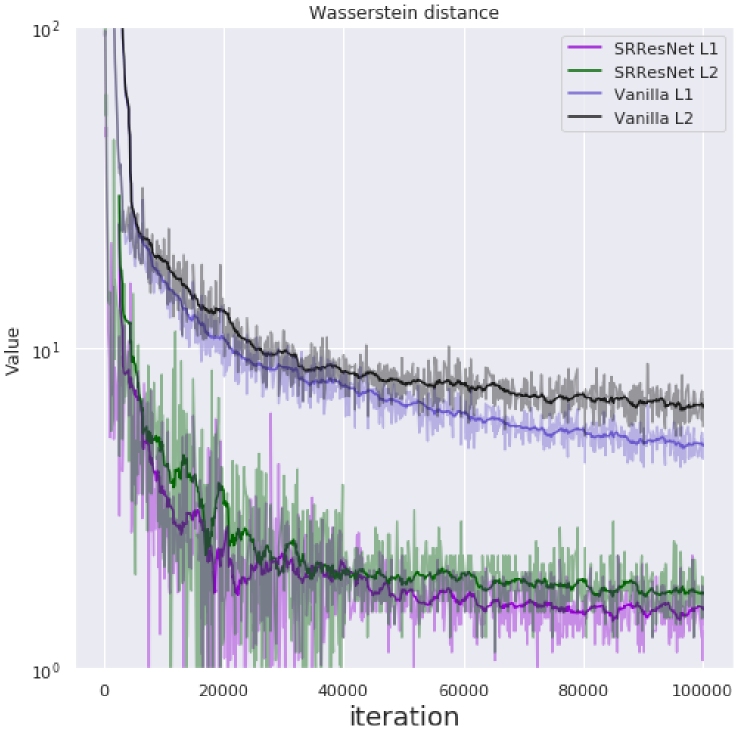


Figure 3.5: Wasserstein distance convergence during training. Both L1 or L2 content loss functions were employed for the CNN baseline and SNGAN models.

3.3.1.2 Training stability

We analyzed the training stability of WGAN-GP and our SNGAN for a variety of learning rates. Fig. 3.6 shows the discriminator loss during training. A substantial spike for WGAN-GP was observed at ∼10k iterations, where the discriminator caused the previously learned weights of the network to reset. When the discriminator failed to learn, the generator fooled the discriminator with trivial samples, regardless of whether the distribution of generated images moved towards the target distribution. Therefore, a stable discriminator was critical for the network to generate samples that resembled the desired reference images. Table 3.2 suggests that our SNGAN discriminator remained stable at a variety of learning rates while WGAN-GP encountered gradient overflow for learning rates greater than $7e-5$.

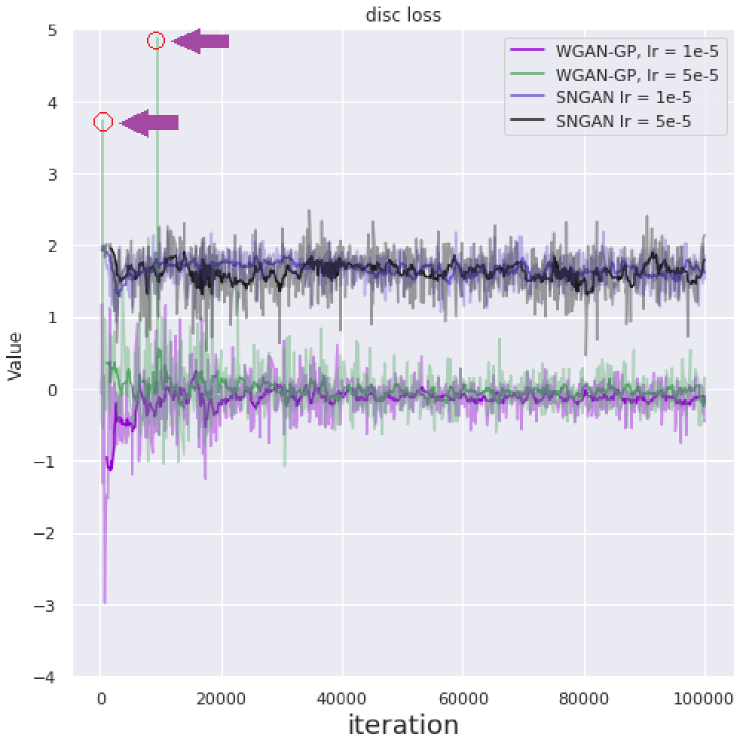


Figure 3.6: Discriminator training loss with different learning rates. WGAN-GP was not stable when the learning rate was set to be $5e-5$. Spikes are highlighted by the purple arrows.

3.3.2 Scenario 2: Multiple CT parameter normalization

Our SNGAN approach achieved better perceptual quality in terms of sharpness and texture compared to the baseline CNN model. Enhancement was more pronounced on the coronal and sagittal plane, as shown in Fig. 3.4. SNGAN was able to recover textures that resembled the reference for all three image conditions displayed.

3.3.2.1 Radiomic feature analysis

While the proposed SNGAN approach yielded superior image perceptual quality, we also examined whether the generated textures have similar statistical characteristics to the reference. Fig. 3.7 depicts the distribution of radiomic feature values after dimensionality reduction using t-SNE. The SNGAN model (green) resulted in transformed radiomics feature values that overlapped with the feature values extracted from the reference image (red). In comparison, the CNN baseline model (orange) pushed the distribution away from the reference (A, B, D, E, G, H). It should be noted that SNGAN was suboptimal in transforming radiomics features for condition B (25% dose, smooth kernel) and I (100% dose, sharp kernel).

Among the four selected features that were used to predict lung nodule malignancy in a previous study [82], scans normalized using SNGAN had smaller absolute errors compared to CNN-generated scans, as illustrated in Fig. 3.9. Suboptimal performance for SNGAN in conditions B and I (Gray-Level-Non-Uniformity and Sum Entropy) was noticeable, underscoring what was observed in the t-SNE analysis. Fig. 3.8 depicts the level of reproducibility for a large set of features. Feature values of unnormalized images had poor agreement for conditions B, C, D, G and H. In most cases, SNGAN was able to mitigate the variability, resulting in moderate to good agreement for these conditions. However, the CNN failed

Table 3.2: SNGAN and WGAN-GP stability. Spectral-norm GAN accepted a larger range of learning rates while WGAN was more sensitive to changes.

lr	SNGAN	WGAN-GP
1e-5	✓	✓
5e-5	✓	not stable
7e-5	✓	✗
1e-4	✓	✗

to effectively mitigate the variability in most conditions. On average, SNGAN increased the number of radiomic features that achieved “good” agreement from 326 to 357 (9.5% increase) and features that achieved “moderate” agreement from 67 to 100 (49% increase). The number of poor agreements decreased from 255 to 191 (25% decrease). In comparison, the baseline CNN model decreased the features with “good” and “moderate” agreement from 326 to 214 and from 67 to 65, respectively. It also increased the poor agreement number from 255 to 369, which was consistent with our observation in t-SNE visualization. SNGAN could not mitigate the variability observed in certain texture features (e.g., GLCM, GLRLM, GLSZM), but these features never achieved “good” agreement under any of the evaluated conditions.

3.3.2.2 Lung nodule detection evaluation

In Table 3.3, after SNGAN normalization, the nodule detection algorithm achieved sensitivity and F1-scores that were more consistent to what was obtained on the reference scan when compared to CNN and unnormalized scans. On average, the CCC increased by 25% and 75% for SNGAN on sensitivity and F1-score metrics, respectively, when compared to unnormalized scans. These results demonstrate that not only SNGAN can mitigate variability in hand-crafted radiomic features, but it also enables CAD algorithms to perform more consistently across a variety of input scans. However, it should also be noted that in condition A, the baseline CNN model clearly outperformed SNGAN (0.7855 versus 0.6431). Condition A was reconstructed using a smooth kernel. CNN-based models output images that are smoother in appearance, which may explain why the CNN achieved results that had a higher level of agreement than SNGAN.

Table 3.3: Level of agreement of nodule detection performance metrics. CCC was used to measure the agreement of performance metrics when evaluating the nodule detection algorithm on unnormalized and normalized (CNN, SNGAN) scans compared to the reference scan. The highest level of agreement for each image condition for sensitivity and F1-score is bolded.

Condition	Sensitivity			F1-score		
	unnormalized	CNN	SNGAN	unnormalized	CNN	SNGAN
A	0.5081	**0.7855**	0.6431	0.3542	**0.7272**	0.6593
B	0.6618	0.7007	**0.7761**	0.3834	0.7112	**0.7900**
C	0.5909	0.7842	**0.8083**	0.4692	0.6652	**0.7714**
D	0.6353	**0.6728**	0.6508	0.4132	**0.5407**	**0.5407**
E	0.6844	0.6584	**0.7007**	0.5682	0.6493	**0.8409**
F	0.6024	0.7693	**0.8567**	0.3919	0.6494	**0.8513**
G	0.5476	0.7422	**0.8144**	0.3909	0.6056	**0.6643**
H	0.5132	0.6348	**0.6904**	0.3381	0.5896	**0.6111**
I	0.6533	0.7508	**0.8302**	0.3556	0.6579	**0.6997**
Average	0.5997	0.7221	**0.7523**	0.4072	0.6440	**0.7143**

3.4 Discussion

Few studies that present a novel normalization technique investigate the impact of the proposed technique on computed radiomic features and downstream clinical tasks. By going beyond image quality metrics, our study addresses a critical gap in understanding the impact of normalization techniques on downstream tasks such as radiomic analysis and CAD. As shown in Fig. 3.8, SNGAN is able to mitigate the effect of variability on most of the first order intensity features. However, both SNGAN and CNN methods are not as effective at mitigating variability in texture features. Nevertheless, SNGAN is able to still achieve moderate agreement in many texture features, much more than the baseline CNN model. Conditions G, H and I are particularly challenging. One factor is that those image conditions are generated using a sharp kernel that increases spatial resolution but results in noisier reconstructed images.

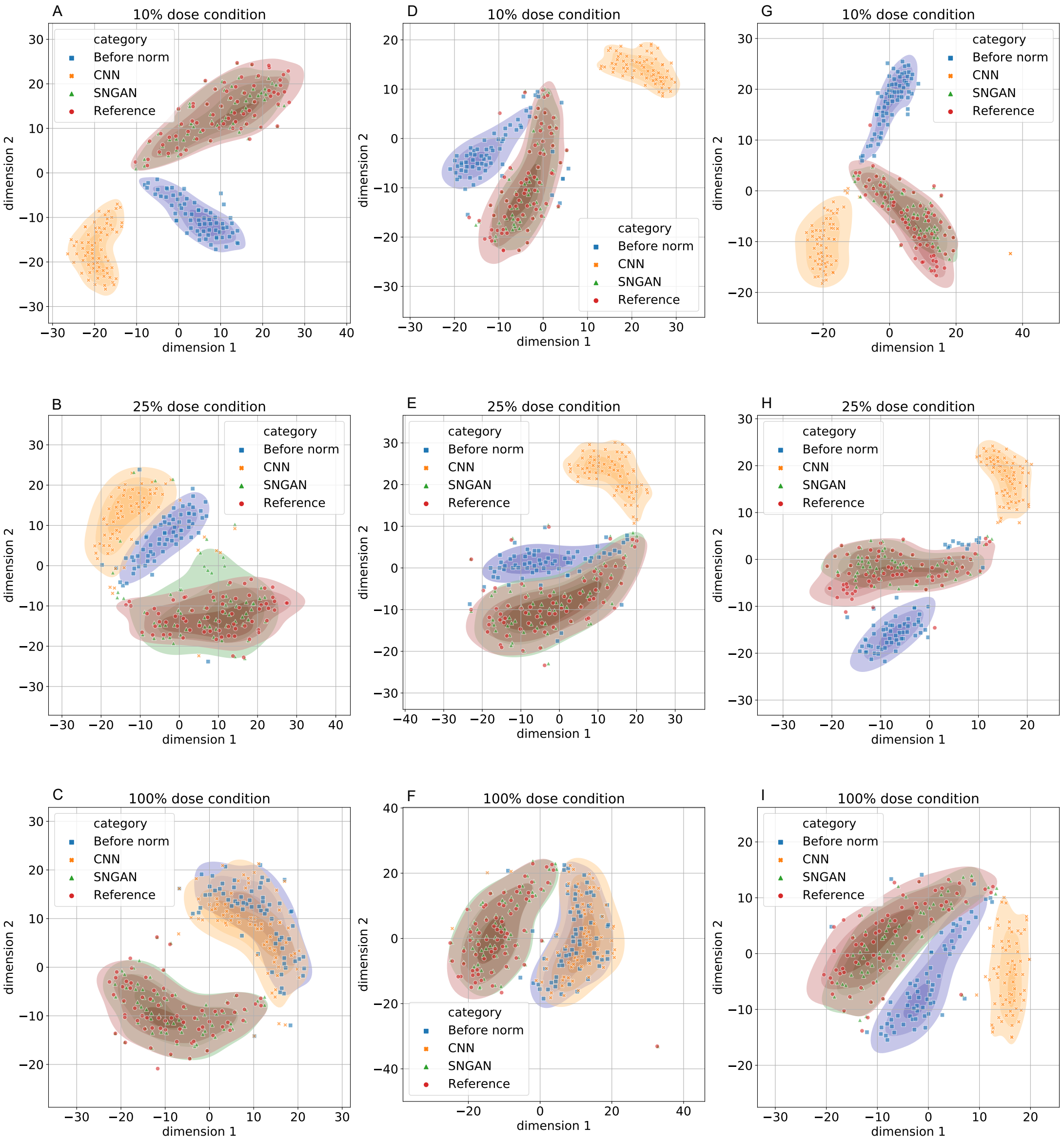
We note that the different evaluations (image quality assessment, radiomic features analysis, nodule detection) provide complementary information. Conclusions drawn from the radiomic features task does not necessarily translate to performance trends observed in the nodule detection task. For example, in Fig. 3.7, we observe that using SNGAN achieves better agreement in radiomic feature values than using a CNN. However, as Table 3.3 shows, compared to CNN, SNGAN achieves an inferior level of agreement in the detection task for condition A. As shown in Fig. 3.8, SNGAN struggles to generate consistent feature values under condition B. Nevertheless, the nodule detection task still performs more consistently using scans normalized using SNGAN compared to the others. Under condition B, SNGAN achieves a better F1-score than C, E, G, D, and I. While some similarities between radiomic and model-learned features may exist, the CAD algorithm likely weighs features differently when performing the final classification. Performing only one of the three analyses would not have given us a comprehensive understanding of the observed trends. The mismatch between radiomic and task performance demonstrates the necessity of a multi-pronged evaluation to provide a more comprehensive view of the benefits and limitations of a normalization method.

There are limitations to our approach. First, as shown in Fig. 3.9, some radiomic features (e.g., Idm under Condition A, SumEntropy under Conditions B & I) perform better (e.g., had smaller absolute errors when compared to the reference scan) when unnormalized than when normalization is applied. One explanation could be that these specific radiomic features are sensitive to the effects of normalization: the CNN overly blurs out texture features while the SNGAN adds too much texture to the images. Second, super-resolution enhancement is an inherently an ill-posed problem with multiple possible solutions. However, the generator of a GAN is deterministic [84]. Stochastic alternatives such as SRFlow [79] may address this issue. Third, we train individual models for each mapping (9 models to cover all 10 image conditions where one condition is designated as the reference). Given that having a model for each mapping is impractical, a conditional GAN (cGAN) could be used to incorporate contextual information (acquisition and reconstruction parameters) during training to learn multiple mappings simultaneously, reusing weights for various input conditions. We focus on the most common types of CT parameter variations (dose, kernel, and slice thickness). It also should be noted that we have not yet explored other sources of intrascanner variability such as pitch, kVp, and detector configurations.

SNGAN was trained and evaluated on data acquired using scanners from a single manufacturer (Siemens Healthineers, Erlangen, Germany). Generalizability of our model remains part of future work, but we have examined a wide range of doses, slice thicknesses, and kernels, which likely overlap with variations seen in scans acquired using scanners from other manufacturers. SNGAN is also potentially generalizable to other imaging modalities (e.g., magnetic resonance imaging, positron emission tomography) whenever paired low and high resolution data are available. Given the difficulty of obtaining paired datasets of the same subject, one avenue of potential exploration is to employ self super-resolution (SSR) algorithms [149].

3.5 Conclusion

Our work addresses the need for techniques to mitigate variability in CT scans due to acquisition and reconstruction parameters. This study presents a 3D GAN-based approach called SNGAN to normalize heterogeneous images using an approach that retains the perceptual characteristics of the reference image. SNGAN is further enhanced using a spectral-norm method to ensure training stability. We evaluate our approach in two scenarios (single CT parameter and multiple CT parameters) and three different experiments (image quality assessment, radiomic feature analysis, task-based evaluation). We show that SNGAN, when compared with other normalization techniques, achieves better image quality metrics, reduces the variability in radiomic feature values, and achieves a higher level of agreement on the nodule detection task.



(a) smooth

(b) medium

(c) sharp

Figure 3.7: Distributions of radiomic features by t-SNE 2D visualization. The distribution of radiomic feature values generated from unnormalized images was clearly different than the distribution of values generated from reference images. SNGAN transformed the distribution such that it overlapped with the reference. Conversely, the baseline CNN model failed to correct for distributional differences and in some cases, made the differences greater.

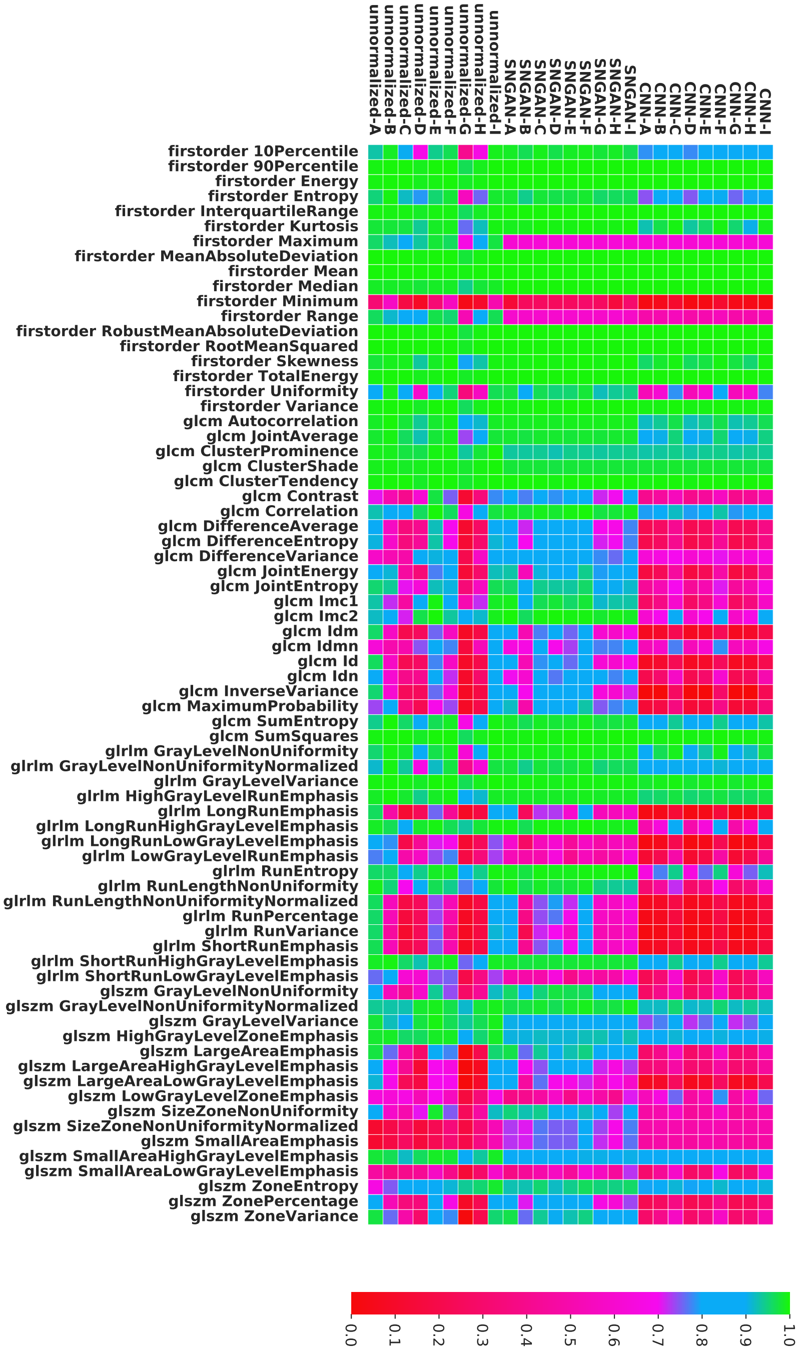


Figure 3.8: Level of agreement of radiomic feature values. CCC was used to measure the agreement for various mapping conditions (e.g. CNN outputs and reference). A CCC of 1.0 corresponds to perfect agreement. A CCC greater than 0.9 was defined to be “good” agreement; a CCC between 0.8 and 0.9 was interpreted as “moderate” agreement; and a CCC less than 0.8 was interpreted as “poor” agreement.

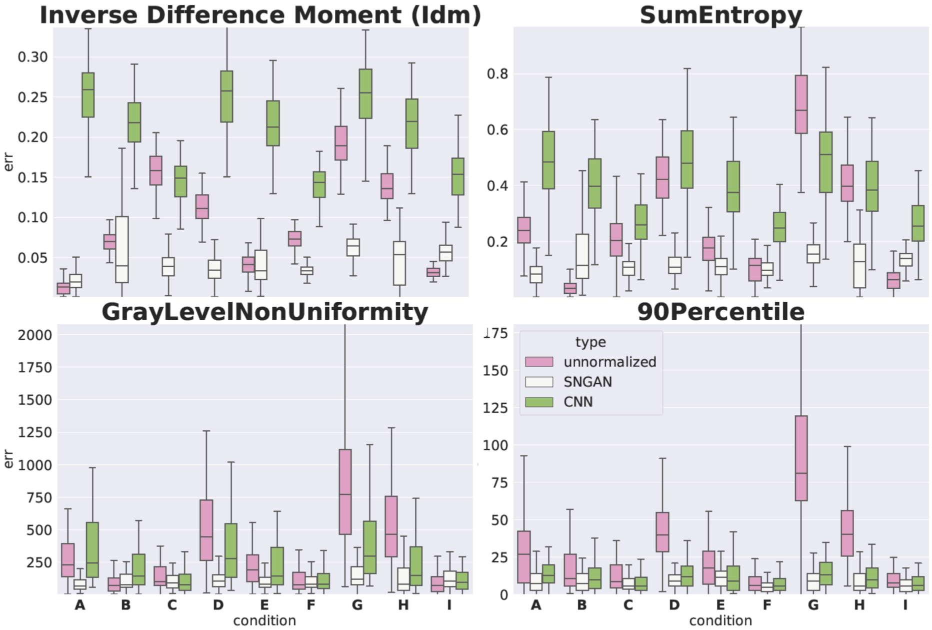


Figure 3.9: Absolute error in radiomic features.

CHAPTER 4

Conditional Normalizing Flows

Mitigating the effects of varying computed tomography acquisition and reconstruction parameters is a challenging inverse problem, where and multiple solutions are plausible. This paper presents CTFlow, a normalizing flow-based method for translating images acquired and reconstructed using different doses and kernels to a reference scan. Unlike existing state-of-the-art image denoising and translation approaches that only generate a single output, flow-based methods learn the explicit conditional density, capture the uncertainty associated with restoration, and output the entire spectrum of plausible solutions. We harness these capabilities to generate more realistic restored reference scans. To evaluate the performance of CTFlow, first, we compare CTFlow with other denoising techniques by training and testing it on the AAPM-Mayo Clinic Low-Dose CT Grand Challenge dataset. CTFlow achieves superior performance for both peak signal-to-noise ratio and perceptual quality metrics. Second, we train and evaluate CTFlow on 186 low-dose CT chest scans from our institution that are reconstructed at different doses and kernels, analyzing the difference in restored reference scans on the performance of a lung nodule detection algorithm. CTFlow produces more consistent predictions across all dose and kernel conditions than the state-of-the-art techniques based on generative adversarial networks (GAN). Third, we investigated generalization by evaluating a pretrained CTFlow model on a publicly available low-dose CT chest dataset. We show that CTFlow maintains higher image fidelity than GAN-based methods. In summary, normalizing flow performs state-of-the-art CT image translation and provides additional information through its ability to quantify restoration uncertainty.

4.1 Introduction

Mitigating effects of variations in computed tomography (CT) images is an important task to restore the high-quality details from the inputs due to non-standard acquisitions or reconstructions. Multiple factors can cause variations, such as acquisitions (dose) and reconstructions (kernels). For example, even small amounts of noise in low-dose CT acquisitions can result in large inconsistencies in clinical evaluations for downstream image analysis tasks such as lung nodule detection and segmentation [28,29]. Therefore, robust models to mitigate the effects of dose variation are critical to ensuring reliable quantitative imaging features from those translated images. In Chapter 3, we explored using a conditional GAN to mitigate effects of variations in CT scans.

However, studies [49,84,152] have shown that cGANs suffer from mode collapse and they are prone to ignore the input noise vector z. All GAN-based methods mentioned in Chapter 2 discourage using the random vector z, and therefore the mapping is deterministic. However, inverse problems such as denoising are ill-posed. These approaches are fundamentally limited in their ability to output the entire spectrum of plausible solutions. This limitation often results in the introduction of artifacts or omission of important anatomical landmarks that may impact computer-aided diagnosis algorithms. A better solution is to explicitly model uncertainty along with image restoration [40]. Interest has grown recently in combining uncertainty with neural network models. Schlemper [106] developed Bayesian inference through Markov chain Monte Carlo (MCMC) variational dropout [32, 57] on a deep cascade of CNNs. Adler [4] used posterior sampling in Bayesian inversion for a conditional WGAN. Both techniques provide robust image restoration from low-quality input data. Tanno et. al [118, 119] created a dual-network architecture that estimates the mean and covariance of the Gaussian conditional distributions on low-resolution input. Using uncertainty modeling, they were able to quantify the risk of generating distortions when performing superresolution of diffusion magnetic resonance imaging (MRI). However, all these works approximate posteriors by variational inference, which can be challenging when dealing with high dimensional distributions such as medical images.

Normalizing flow algorithms, which compute the exact posterior density directly by optimizing likelihood, overcome these limitations [24,25]. These generative models have shown success in conditional image generation for natural images [79]. To the best of our knowledge, there has only been a single attempt to apply this method to medical applications. Denker [23] employed a normalizing flow model conditioned on LDCT reconstruction by Filtered Back-Projection (FBP) to improve reconstruction quality from raw sinogram data. It was unclear, however, how it could be extended to directly denoise LDCT scans in the image space. As such, this paper presents CTFlow, an approach inspired by [56,79], which aims to solve inverse problems and mitigate the variations in CT via maximizing the explicit likelihood of a standard CT scan given a non-standard one. Normalizing flow has two important advantages: 1) The translated low-dose CTs have minimal artifacts because the output is a maximum likelihood estimate that closely matches the target reference distribution; and 2) unlike GANs that are susceptible to mode collapse, CTFlow is able to explore multiple solutions to reduce inference uncertainty. We demonstrate how these advantages provide consistent computer-aided diagnosis (CAD) performance when characterizing the same imaging abnormality across a variety of input conditions in the context of lung nodule detection.

The contributions of our work are as follows:

- The conditional normalizing flow was applied to a low-dose CT denoising task and outperformed state-of-the-art methods in terms of both image fidelity and perceptual quality on the AAPM-Mayo Clinic Low Dose CT Grand Challenge dataset.
- The conditional normalizing flow was applied to the task for mitigating both dose and kernel variations on our in-house UCLA dataset, showing that the approach achieves more consistent downstream lung nodule detection results compared to GAN-based

methods.

- We present a novel autoencoding technique applied to manipulating the latent space, resulting in better generalization for external datasets.
- The proposed method allows for the measurement of uncertainty using diverse output images that are obtained by sampling the latent space.

4.2 Method

4.2.1 Conditional Normalizing Flow

A deterministic approach to image translation, such as using a CNN, finds a mapping function $y = g_\theta(x)$ that takes a non-standard input image x and outputs an image y under reference condition. For example, x could be a low-dose CT image and y could be a normal-dose (routine-dose) image. For the purpose of illustration, we use this denoising narrative in the following method section. However, note that x could refer to any CT image from a non-standard protocol with multiple variations in parameters and y can be a predefined reference condition. Flow-based image translation aims to approximate the density function $\Pi_{y|x}(y|x, \theta)$ using maximum likelihood estimation. Normalizing flow gradually transforms a simple initial (Gaussian) density function $p_z(z)$ to a target distribution $\Pi(y|x)$ using an invertible neural network $y = g_\theta(z; x) \leftrightarrow z = g_\theta^{-1}(y; x) = f_\theta(y; x)$, where g and f are the decoding and encoding functions. By the change of variables theorem, we have

$$\Pi_{y|x}(y|x, \theta) = p_z(z)\left|\det \frac{\mathrm{d}z}{\mathrm{d}y}\right| = p(f_\theta(y; x))\left|\det \frac{\mathrm{d}f_\theta(y; x)}{\mathrm{d}y}\right|, \tag{4.1}$$

which can be trained by maximizing the log-likelihood. In practice, a multilayer flow operation is preferred because single-layer flow is not able to identify complex non-linear relationships within data. We decompose f_θ into a series of invertible neural network layers $h^n, n = 1, 2, 3..., N$. $h^n = f_\theta^n(h^{n-1}; e(x))$, where we used a deep CNN $e(x)$ to extract salient

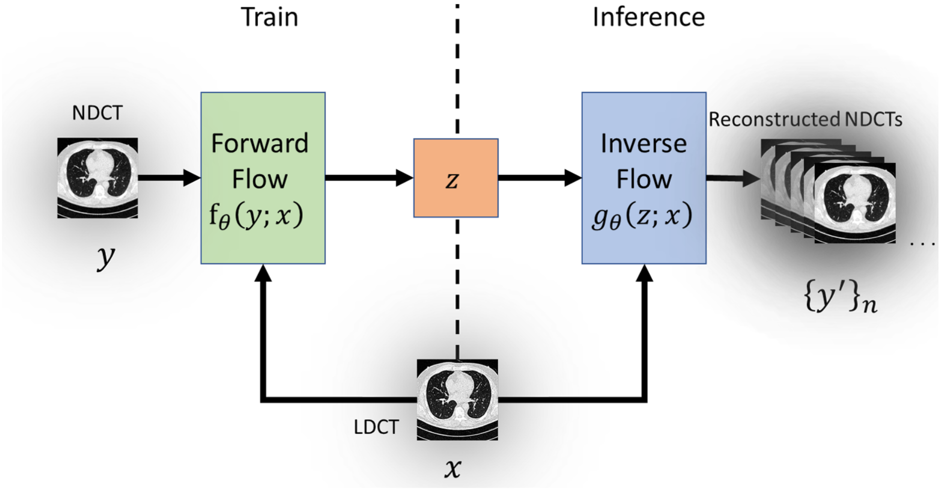


Figure 4.1: Framework of forward flow and inverse flow. LDCT: low-dose computed tomography; NDCT: normal-dose (or routine-dose) computed tomography.

feature maps of input x to condition on flow layers. For an N-layer flow model, $y = h^0$ and $z = h^N$. Therefore, applying the chain rule, we aim to maximize log-likelihood as shown in Equation 4.2. The first term is tractable since it is a Gaussian. In addition, we only need to calculate the determinant of the Jacobian $\frac{\mathrm{d}f_\theta^n}{\mathrm{d}h^{n-1}}$ for each flow layer in this formulation. We note that the second term requires special attention to be executed efficiently, as discussed in 4.2.2.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \log p_z(z) + \sum_{n=1}^{N} \log \left| \det \frac{\mathrm{d}f_\theta^n(h^{n-1}; e(x))}{\mathrm{d}h^{n-1}} \right| \tag{4.2}$$

Once the training is complete, we apply the decoding function $g_\theta(z;x)$ with random latent variables z from an independent and identically distributed Gaussian. The use of latent variables z allows us to explore various restored images y' conditioning on the same non-standard input x. The framework of CTFlow is illustrated in Figure 4.1.

4.2.2 Flow layers

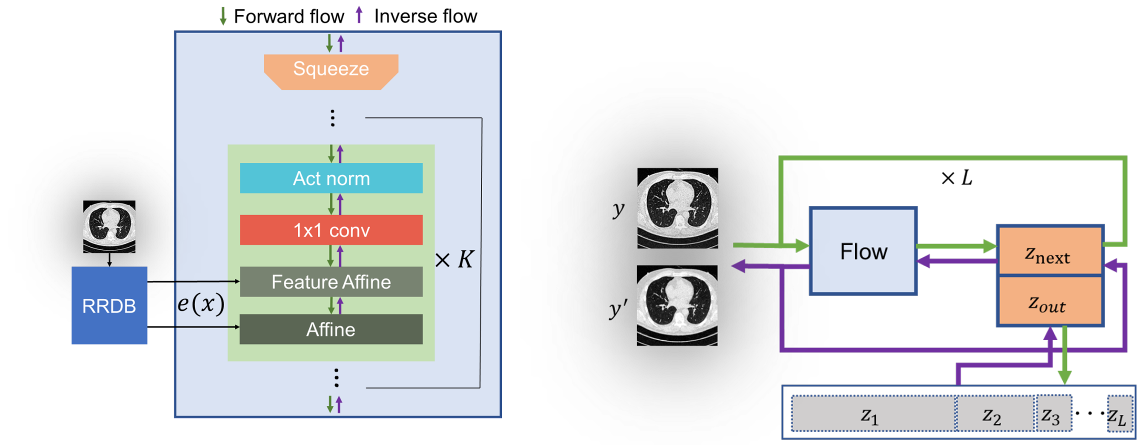
As discussed, flow layers must meet two requirements: 1) be invertible 2) be tractable Jacobian determinant. We follow the triangulation trick developed by NICE [24]. The core idea is to use affine coupling layers adopted with a conditional variable, noted as the self-conditional affine layer. We first equally split the channels into $(h_1^n, h_2^n) = \texttt{split}(h^n)$, and apply affine transformation on h_2 while keeping an identity transform on h_1, as illustrate in Equation 4.3. `NN` is a shallow convolutional neural network that is used to compute scale and shift factor in spatial coordinates i, j . Thus, by definition, Jacobian of h^{n+1} is a lower triangular matrix. The log determinant is simply $\texttt{sum}(|s|)$.

$$\begin{aligned} h_1^{n+1} &= h_1^n \\ (s, t) &= \texttt{NN}_\theta^n(h_1^n; e(x)) \\ h_2^{n+1} &= \exp(s) \odot h_2^n + t \\ h^{n+1} &= \texttt{concat}(h_1^{n+1}, h_2^{n+1}) \end{aligned} \tag{4.3}$$

Activation normalization : Channel-wise batch normalization [48] that per output channel output has zero mean and unit variance.

Invertible 1x1 conv: In contrast to RealNVP which shuffles the channel order before affine coupling split, we follow the study in [56], utilizing a learnable 1x1 convolution $h_{ij}^n = W h_{ij}^{n-1}$, where W is a square matrix with dimension $c \times c$ (channels). Each spatial element ij in h is multiplied by this 1x1 convolution matrix W. The log determinant is $hw \log \texttt{sum}(\det(W))$ and can be computed efficiently using PLU factorization as suggested in [56]. Moreover, the inverse 1x1 operation is trivial to compute because the cost of calculating the inverse matrix W^{-1} is relatively small.

Feature conditional affine: We already have a self-conditional layer in the conditioning setting that partially incorporates the noisy image feature maps into the flow steps. Here, we aim to impose a more vital interaction between feature maps extracted $e(x)$ and activation maps h. To achieve this, as equation 4.4 shows, we directly compute the scale and shift



(a) Flow module. RRDB: Residual in Residual Dense Blocks

(b) Multiscale architecture

Figure 4.2: Flow module and multiscale CTFlow architecture. $K = 16, L = 3$ were used in this work as suggested in [79].

factor from $e(x)$. Overall, the basic network structure of flow is shown in Figure 4.2a.

$$\begin{aligned} (s, t) &= \mathtt{NN}_\theta^n(e(x)) \\ h^{n+1} &= \exp(s) \odot h^n + t \end{aligned} \tag{4.4}$$

4.2.3 Multiscale architecture

Since flow is inherently invertible, it requires input x and latent space vector z to have the same dimension. However, in most cases $\Pi_{y|x}(y|x, \theta)$ is a low-dimensional manifold on a high-dimensional input space. Significant computational resources are wasted when the flow model is imposed with dimensionality higher than the dimension of true latent space. As a result of a multiscale architecture in RealNVP, we simplify the model and improve the estimation of $\Pi_{y|x}(y|x, \theta)$ at multiple levels. The overall multiscale architecture is depicted in Figure 4.2b, where we equally divide each output z into (z_{out}, z_{next}), while recursively feeding z_{next} to the next level and, before directly outputting z_{out} for maximum log-likelihood estimation at the end.

4.2.4 Training details

We used a clinical dataset "AAPM-Mayo Clinic Low-Dose CT Grand Challenge" by Mayo Clinic to train and validate our model for image quality in CT denoising task. The dataset consists of 5,936 abdominal CT images at 1.0 mm slice thickness taken from both routine-dose and simulated quarter-dose pairs from 10 patients. Among them, 80% were used for training and 20% were reserved for validation. All images were randomly cropped into patches of 128×128 pixels, excluding the area that was mostly air. Our comparison was conducted using GAN-based approaches (GAN, WGAN-MSE, WGAN-VGG, and SNGAN [133, 136, 139]), CNN-based approach (SRResNet) [73] and a denoising algorithm based on collaborative filtering, Block-matching and 3D filtering (BM3D) [22] for comparison.

CTFlow training is divided into two parts.

- Our first step was to train a CNN based on Residual-in-Residual Dense Blocks (RRDB) [129], which has been extensively explored in many superresolution works. This CNN contains 14 RRDB blocks and serves as our feature extractor for low-dose images. The RRDB network was trained using $\mathcal{L}_1$ loss for 60k iterations. The batch size was 16 and the learning rate was set to 2e-4. The Adam optimizer was used with $\beta_1, \beta_2 = 0.9, 0.99$. After training, all layers of RRDB were frozen and used only for feature extraction. Feature maps were derived from {2, 6, 10, 14} block outputs. Afterward, the outputs of each block were concatenated into a conditional feature map $e(x)$.

- In the same manner, we trained the CTFlow model with a batch size of 16 and 50k iterations. The learning rate was set to 1e-4 and halved at 50% 75% 90% and 95% of the total training steps. Negative log-likelihood (NLL) loss was used, and the network took 3 days to train on an NVIDIA RTX 8000 GPU. The peak GPU memory usage was 39GB. An example of training NLL curve is shown in Figure 4.3. Unlike the adversarial training of GANs that requires two loss functions, our network had only one loss and was easily optimized. NLL was stable and decreased monotonically.

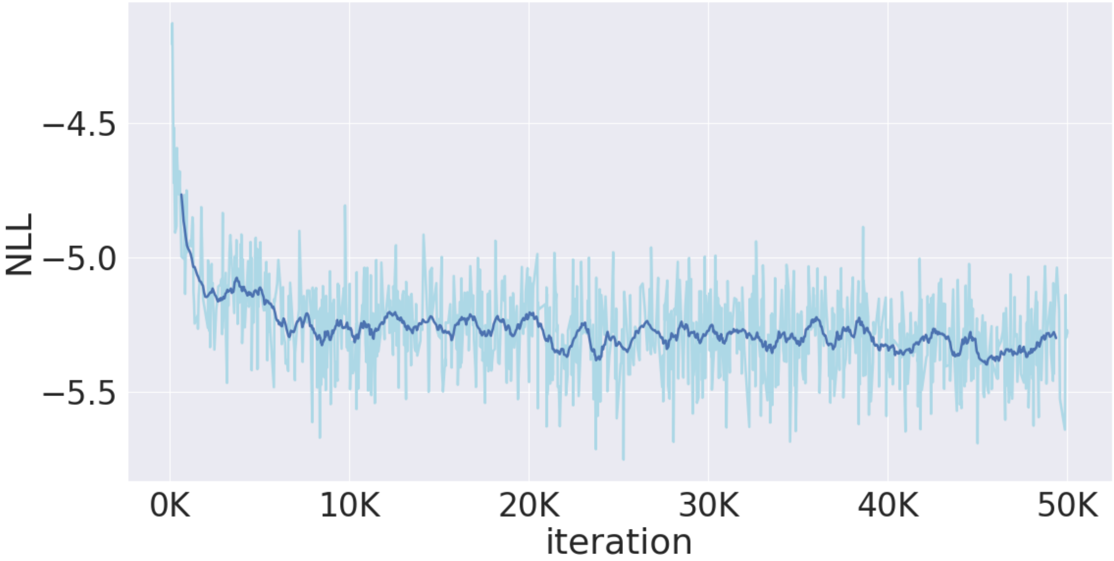


Figure 4.3: Learning curve for negative log-likelihood (NLL) for AAPM dataset.

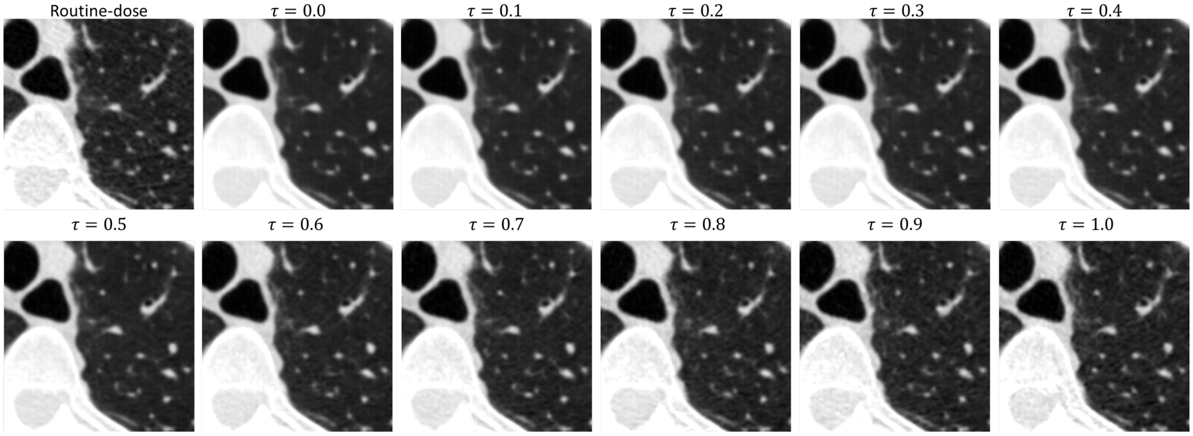


Figure 4.4: Denoised images in AAPM dataset with different temperature τ settings. The texture emerges in lung parenchyma as τ increases.

4.2.5 Autoencoding

Deep learning techniques for medical imaging are plagued by the problem of model generalization. Usually, due to the intra- and inter-variability of CT acquisition and reconstruction protocol, a model trained on one dataset will not generalize well to another dataset from a different institution. We observed suboptimal performance when applying our model to the public Mayo clinic low-dose CT dataset [87] using a direct mapping $g_\theta(z; x)$, which was trained on the UCLA dataset. GAN showed a similar trend in a previous study [69]. Using the powerful representation of the latent space vector z, we propose to develop a new autoencoding technique to address this problem. On the basis of this, we developed a more sophisticated image denoising technique by using latent space normalization, as described in [79]. The autoencoding technique is useful when CTFlow is applied to external datasets (e.g., CT scans acquired using different protocols and hardware platforms). This procedure is carried out as first encoding the input LDCT image by conditioning on itself as $\hat{z} = f_\theta(x; x)$. Like the auto-encoder, this encoding vector contains the latent code for reconstructing a clean version of x, or $g_\theta(\hat{z}; x)$. However, since we condition on x, which has a slightly different image appearance than an unknown ground truth y, the resulting $\hat{z}$ does not follow a standard Gaussian distribution, while g_θ expects $z \sim N(0, I)$ by feeding f_θ with $(y; x)$ pairs. If we are to make our assumption valid, we must normalize its statistics to $N(0, \tau)$ where τ is a temperature scaling term that dictates diversity of outputs. Suppose a collection of $\{z\}_N$ comes from $N(0, \tau)$, by definition, we have its empirical mean and variance as

$$\mu \sim \mathcal{N}(0, \frac{\tau}{N}), \quad \sigma^2 \sim \Gamma(\frac{N-1}{2}, \frac{2\tau}{N-1}). \tag{4.5}$$

Meanwhile, we also sample a collection of $\{\hat{z}\}_N$ by autoencoding, and compute the mean $\hat{\mu}$ and variance $\hat{\sigma}^2$. In this work, we normalize the statistics spatially for each latent vector channel, because contrast shift and global noise characteristics are the primary sources of discrepancy between two different CT datasets. The latent vector normalization is therefore formulated as

$$z_c = \frac{\sigma_c}{\hat{\sigma}_c}(\hat{z}_c - \hat{\mu}_c) + \mu_c, \forall c \in C, \text{C: set of channels} \tag{4.6}$$

We compared the z latent vector normalization results with direct CTFlow inference and SNGAN in 4.4.2.

4.3 Experiments

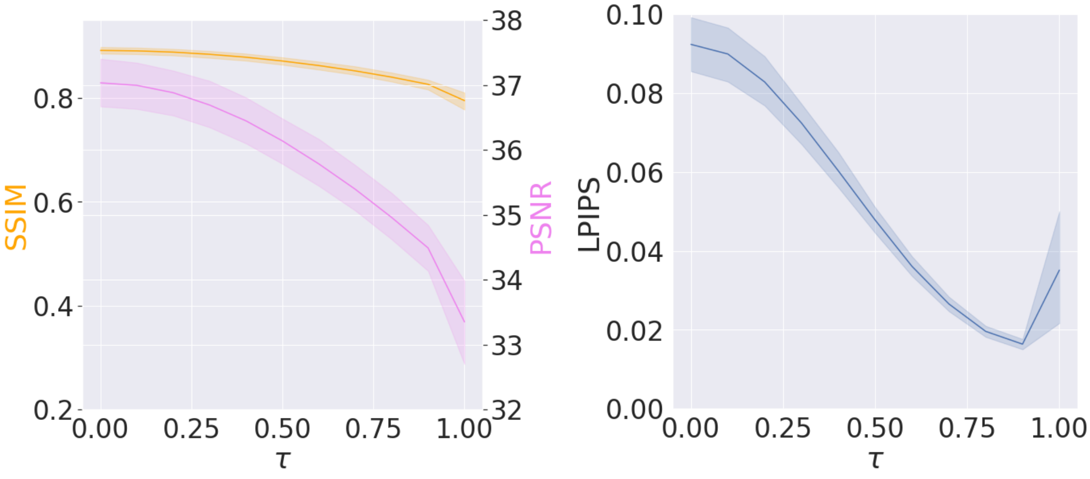
We perform the following experiments to evaluate our CTFlow approach: 1. We assessed image quality and compare with other previously published low-dose CT denoising techniques using the AAPM dataset and 2. We evaluated the ability of CTFlow to mitigate different sources of variations(such as dose and kernels) towards achieving more consistent performance of a computer-aided lung nodule detection algorithm.

4.3.1 AAPM dataset evaluation

This section assesses CTFlow's performance in image quality in comparison with other state-of-the-art solutions. We compute image quality metrics using the peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) [147].

4.3.1.1 Texture control

Temperature (τ) has a direct impact on the output image textures. The higher the temperature, the more vivid the images and textures are; and the lower the temperature, the more blurry the images. We observed the same trend in the CT dataset in Figure 4.4. In the extreme case where $\tau = 0.0$, $p_z(z)$ collapses into the Dirac Delta distribution and the mapping becomes fully deterministic. Flow-based models achieve the best performance when sampled slightly less than one, as empirical evidence has demonstrated [56]. In order to iso-



(a) Image fidelity

(b) Perceptual quality (LPIPS)

Figure 4.5: Trade-off between image fidelity (distortion) and perceptual quality. Temperature settings enable quantitative adjustment to distortion and perceptual appearances. LPIPS attains its minimum near 0.8.

late the setting for the best image quality, we conducted a parameter sweep on τ. Results are presented in Figure 4.5. A trade-off between perceptual quality (as measured by LPIPS) and image fidelity (as measured by PSNR and SSIM) was observed. As a result, we chose τ to be 0.8 to ensure reasonable fidelity while achieving the best perceptual quality.

4.3.1.2 Image quality assessment

In Table 4.1, we provide the results for image quality metrics, while Figure 4.6 presents examples of denoised CT images. We aimed to achieve the best perceptual quality (lowest LPIPS), while maintaining a higher degree of fidelity (PSNR, SSIM). In Figure 4.6, while both BM3D and SRResNet generated the highest PSNR, the resulting images were overly smoothed and lacked high-frequency components. Important texture details were lost in the restoration, which could negatively impact detection performance of a radiologist or CAD that relies on texture features to characterize lesions. Furthermore, CTFlow achieves 6%

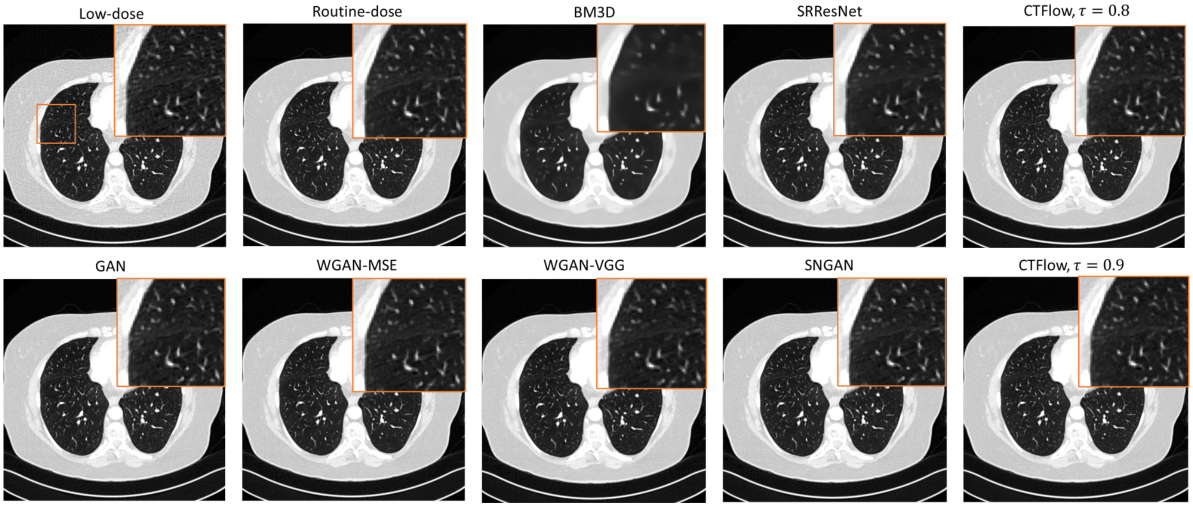


Figure 4.6: AAPM dataset denoising results comparison. Streaking artifacts around lung fissures are observed for all GAN-based approaches. CTFlow has no such distortion.

better perceptual quality compared to SNGAN, while maintaining the best PSNR fidelity in comparison with GAN-based approaches (+0.13 dB for WGAN-VGG). The GAN-based methods suffer from low-dose CT artifacts that are made of bright and dark streaks in the direction of more significant attenuation [13]. Images produced by CTFlow, however, did not have such artifacts.

4.3.2 Nodule detection evaluation

CTFlow has the potential to generate fewer artifacts and to be more consistent with the input. This section evaluates CTFlow's performance of mitigating multiple CT variations with respect to lung nodule detection tasks. Here, we use a dataset of 186 chest LDCT exams collected at our institution that were acquired at an equivalent dose of 2mGy. The raw projection data of scans performed on Siemens CT scanners was exported. Poisson noise was introduced into the raw projection data, as described [44] at levels that were equivalent to 10% of the original dose. Original routine-dose and reduced dose projection data were then reconstructed into an image size of 512 × 512 using three reconstruction

Table 4.1: Generated denoised image quality results on validation images. Red: the best, blue: the second best in that metric category.

	↑ PSNR	↑ SSIM	↓ LPIPS
BM3D	32.81	0.8471	0.1754
SRResNet	31.89	0.8907	0.0865
GAN	30.76	0.8577	0.0379
WGAN-MSE	31.32	0.8636	0.0357
WGAN-VGG	31.37	0.8688	0.0353
SNGAN	31.28	0.8648	0.0345
CTFlow	31.50	0.8631	0.0324

kernels (smooth, medium, sharp) at 1.0mm slice thickness. The dataset was split as follows: 80 scans for training, 20 scans for validation, and 86 scans for testing. In the test set, 43 scans (50%) contained a total of 68 lung nodules. The centroids of these nodules were marked by a trained image analyst using the original radiologist report as a reference. In this study, the reference ground truth condition was 100% dose, medium kernel, and 1.0 mm slice thickness. Our choice of slice thickness and kernel was based on the parameters that are currently recommended for lung cancer screening. We trained three separate models that map low-dose images reconstructed from smooth, medium, and sharp kernels to the ground truth. In an ideal world, based on the fact that all three of these models are mapped to the same reference image condition, the results of lung nodule detection should be consistent across these three models. Note that for each image, we generate 100 samples from a random noise vector for CTFlow. Meanwhile, we also trained a SNGAN model on the UCLA dataset for comparison. In this inverse problem, not only is the dose mitigated, but also the kernel (sharp, smooth to medium). As a result, it is more challenging than a simple denoising problem.

Our CAD system is a nodule detection network adapted from the RetinaNet model [75], a composite model comprised of a backbone network called feature pyramid net (FPN) and two subnetworks responsible for object classification with bounding box regression. The model was trained and validated on the LIDC-IDRI dataset [10], a public de-identified dataset of diagnostic and low-dose CT scans with annotations from four experienced thoracic radiologists. As part of the training process, we only considered nodules that at least three radiologists annotated. A total of 7,607 slices (with 4,234 nodule annotations) were used for training and 2,323 slices (with 1,454 nodule annotations) for validation. A bounding box was then created around the union of all the annotator contours to serve as the reference for the detection model. After training for 200 epochs with Focal loss and Adam optimizer, the model achieves an average precision (AP@0.5) of 0.62 on the validation set.

The intent of our study is not to achieve the highest lung nodule detection performance,

Table 4.2: Analysis of lung nodule detection consistency measured by CCC scores for three pairwise kernel combinations.

	smooth-medium	medium-sharp	smooth-sharp	mean
CTFlow	0.9906	0.9973	0.9914	0.9931
SNGAN	0.8529	0.8438	0.9413	0.8793

but rather to generate consistent prediction across all mappings. After obtaining subject-level sensitivity and precision, we calculate subject-level F1-score by $\mathrm{F1}_i = 2 \times \mathrm{precision}_i \times \mathrm{sensitivity}_i / (\mathrm{precision}_i + \mathrm{sensitivity}_i)$. Since CTFlow has 100 samples for the same input images, we take the average F1 score. Next, we use the Concordance Correlation Coefficient [62] for measuring the consistency of prediction across all three kernels. McBride [86] suggests the following guidelines for interpreting Lin's concordance correlation coefficient. Poor: $<$ 0.9; moderate: 0.90 to 0.95; substantial: 0.95 to 0.99; perfect: $>$ 0.99 and above. We computed pairwise CCC for the smooth, medium, and sharp kernel. Table 4.2 provides the results. CTFlow significantly outperformed SNGAN. The uncertainty caused by training individual mappings diminishes as the sample size increases. The results showed CTFlow's inherent advantages of being able to explore the entire output sample space.

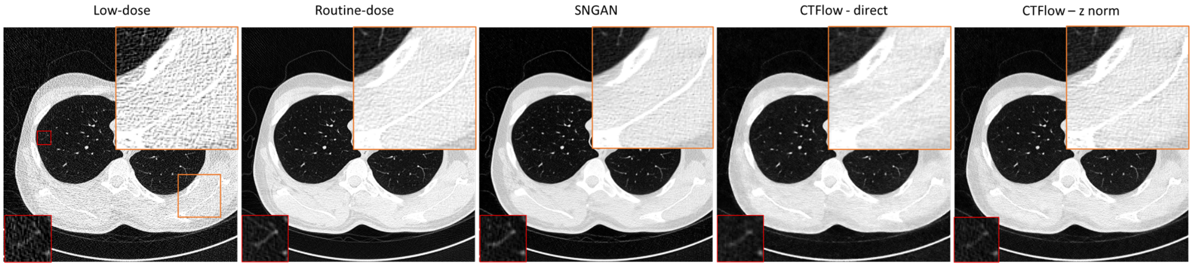


Figure 4.7: Restoration generalization on Mayo clinic low-dose CT image dataset. Pepper and salt artifact can be seen in SNGAN results in the highlighted orange rectangle. It can be seen from in red rectangle that SNGAN results in greater distortion of blood vessels, whereas CTFlow guarantees structural consistency to the low-dose image.

4.4 Applications

We will explore two topics in this section that relate to some practical application of CTFlow: 1). Predict model uncertainty when reconstructing a low-dose image to the routine-dose 2). Improve model performance on an external dataset acquired using protocols that differ from the training dataset.

4.4.1 Restoration uncertainty

Since CTFlow provides a distribution instead of giving a single prediction, we are able to examine the uncertainty resulting from noise when translating a low-dose to a routine-dose scan.

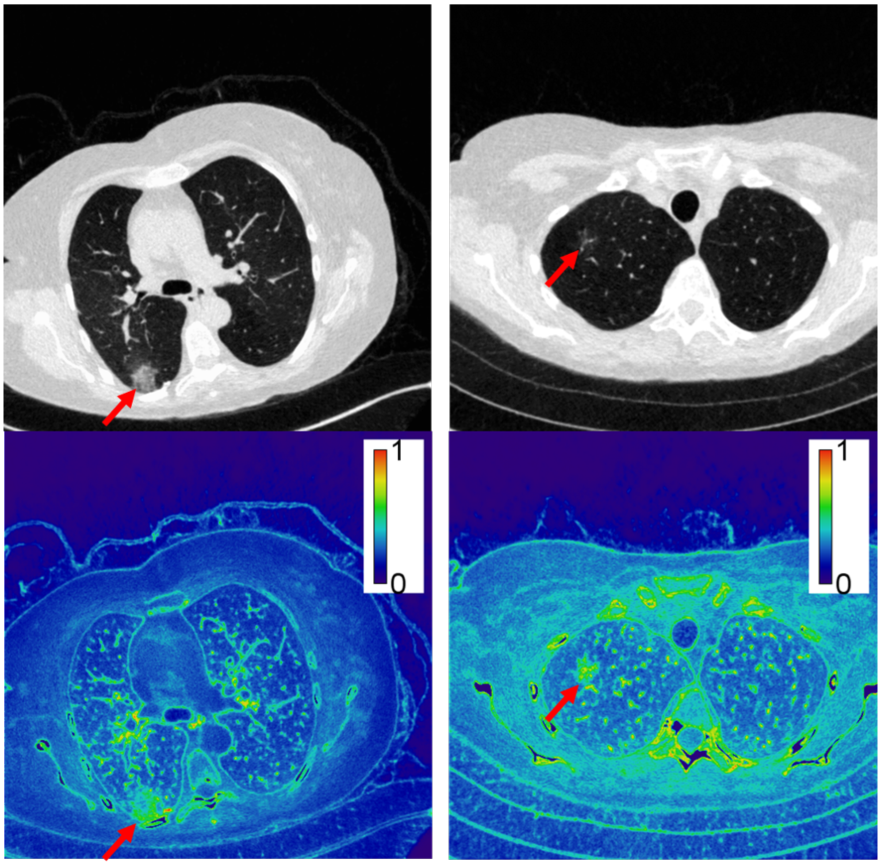


Figure 4.8: Restoration uncertainty characterized by a heat map of standard deviation.

In Figure 4.8, we plot the heat map of standard deviation for the outputs. Based on the arrow displayed above, the area surrounding the nodule has a relatively high intensity level, which indicates that the model has low confidence for restoration in that region since the pixel-wise variation is large. Any other state-of-the-art methods do not provide this information. For ill-posed inverse problems such as denoising, a single prediction is not reliable since the uncertainty of the model cannot be predicted. CTFlow is provided with an opportunity to visualize uncertainty during restoration, allowing us to increase our trust in a denoising model.

Table 4.3: Model generalization performance measured by image quality metrics. Both GAN and Flow were trained on the UCLA dataset but with validation on Mayo clinic low-dose CT image dataset.

	PSNR	SSIM	LPIPS
SNGAN	24.59	0.5160	0.1042
CTFlow direct	24.78	0.5031	0.1335
CTFlow z norm	24.62	0.5349	0.1099

4.4.2 Model generalization

We selected 50 chest CT scans obtained from Siemens scanners in Mayo clinic low-dose CT image and projection data [87]. Standard clinical protocols were followed to obtain CT scans of the anatomical region of interest using routine-dose levels specified by the institution that acquired the data. Poisson noise was then added to the projection dataset to create a simulated lower doses. Low-dose chest scans are provided at 10% of the routine-dose. By leveraging the technique of autoencoding, CTFlow can handle images sampled from a different distribution. Table 4.3 summarizes the results. For natural images, it has been established that GAN-based approaches tend to produce more photorealistic results, which more than offsets its disadvantage of introducing too many distortions. The problem of distortion arises when dealing with data that are far from the training set. However, it is crucial to work with medical images that are consistent with the original input anatomy features. Based on this premise, the primary objective is to maintain image fidelity and avoid distortion. The latent space normalization implemented by CTFlow ensures a consistent anatomical similarity for the input image, and at the same time improves perceptual quality significantly. The SNGAN, on the other hand, is inferior in terms of image fidelity as measured by PSNR and SSIM, which indicates more distortions or artifacts in the restored routine-dose image. The distortions can be seen in Figure 4.7, whereas the CTFlow results do not exhibit such artifacts.

4.5 Discussion and conclusion

We developed a conditional normalizing flow model, CTFlow, to mitigate the effect of CT variations seen in non-standard CT scan inputs, which led to improved lung nodule detection performance over a GAN-based approach. In addition, we demonstrated that CTFlow could learn more accurate data distributions by learning the explicit likelihood. With the AAPM dataset evaluation, we found superior image quality without sacrificing perceptual quality. Moreover, the flow model provides a measure of uncertainty in restored images that CAD algorithms can leverage to identify regions more susceptible to noise and artifacts during the restoration process. However, we should also acknowledge that CTFlow has its limitations too. It is not yet able to achieve the best image fidelity in terms of SSIM score. Furthermore, the user must provide an appropriate temperature for the image generation, which is another hyperparameter to be considered. Figure 4.5 illustrates how too high a temperature setting can lead to undesirable behavior (poor perceptual quality). The alternative state-of-the-art GAN-based works are all based on adversarial training. The advantage of normalizing flows over conditional generative models is that they can offer exact and efficient likelihood computation and diversity of data generation closer to the true distribution. In contrast, the learning objective of GAN does not involve an explicit likelihood function, but rather focuses on generating the best samples. However, this makes the quantitative evaluation of the conditional GAN model biased. Currently prevalent evaluation criteria based on image quality metrics (PSNR, SSIM, etc..) do not address this issue since it is possible to generate realistic samples by memorizing the training data, or missing diversity of the distribution, but still achieve the best quality. Therefore, we suggest task-based evaluations such as nodule detection, in which uncertainty plays a significant role. In GAN, the missing diversity is referred to as mode collapse. Essentially, if the prior is defined over a support that is smaller than the true dimension of the data, which is usually the case for GAN-based models, the likelihood is ill-defined. The high dimensionality required by the latent space vector of normalizing flows is the exact reason why likelihood estimation is successful in our experiments. The distinct differences and relations have been analyzed in [37].

Although it is outside the scope of this work, we should be aware that the general idea of modeling uncertainty offered by normalizing flows can also be extended to other image-processing tasks beyond prepossessing images, such as segmentation, detection, and classification. An example is Chan et.al [15] who applied an approximate Bayesian inference scheme based on posterior regularization to improve uncertainty quantification on covariate-shifted data sets, resulting in improved prognostic models for prostate cancer. Similarly, [104] presented methods to transform pixel-wise uncertainty into structure-wise uncertainty metrics for better brain segmentation, demonstrating their effectiveness in performing more reliable group analysis. A recent preliminary study [96] shows the promise of normalizing flow to detect abnormalities on patches of histopathology images. As a conclusion, the use of a flow-based approach in medical applications requiring both the estimation of density and the generation of samples is a promising direction for the future.

CHAPTER 5

Normalization model robustness

In this chapter, two concepts are discussed to improve the robustness of normalization models. The purpose of Chapter 5.1 is to discuss a novel spatial-temporal convolution technique for improving computation efficiency and analyzing the impact of enhanced computation efficiency on radiomic features. In Chapter 5.2, we describe a holistic approach to understanding the generalizability of different normalization methods, which provides a deeper understanding of the relationship between image metrics, quantitative imaging features, and task-driven evaluation to serve as polestars when constructing a normalization model.

5.1 Spatial-Temporal convolution

While deep-learning-based imaging denoising techniques can improve the quality of low-dose computed tomography (CT) scans, repetitive 3D convolution operations cost significant computation resources and time. We present an efficient and accurate spatial-temporal convolution method to accelerate an existing denoising network based on the SRResNet. We trained and evaluated our model on our dataset containing 184 low-dose chest CT scans. We compared the performance of the proposed spatial-temporal convolution network to the SRResNet with full 3D convolutional layers. Using 8-bit quantization, we demonstrated a 7-fold speed-up during inference. Using lung nodule characterization as a driving task, we analyzed the impact on image quality and radiomic features. Our results show that our method achieves better perceptual quality, and the outputs are consistent with the SRResNet baseline outputs for some radiomics features (31 out of 57 total features). These observations together demonstrate that the proposed spatial-temporal method can be potentially useful for clinical applications where the computational resource is limited.

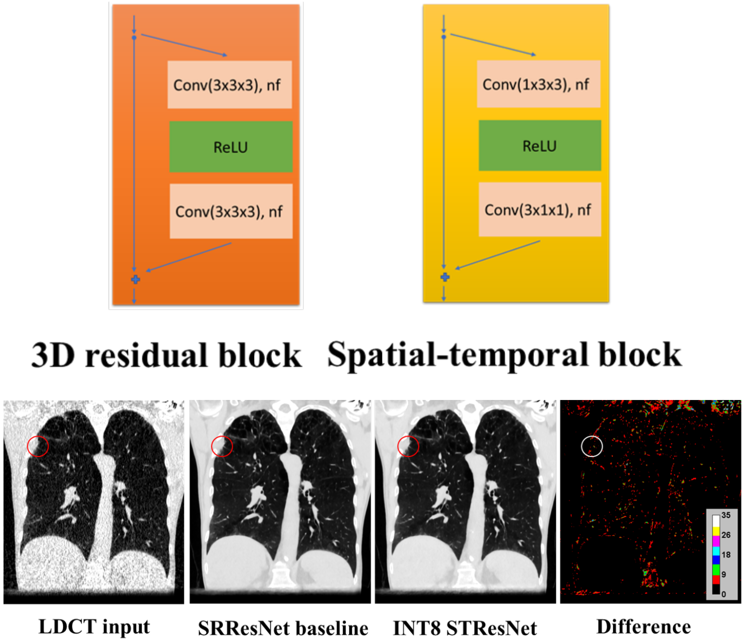
5.1.1 Introduction

Computed tomography (CT) scans provide a detailed characterization of chest anatomy for radiologists to identify lesions in the lung. However, in practice, CT acquisitions are not standardized. Given that higher radiation exposure comes with the risk of harmful radiation, the trend has been to acquire lower dose images at the cost of noisier images. Recent developments in deep learning-based image denoising have yielded a number of approaches to recover high-resolution details from lower resolution inputs. Prior studies have also demonstrated that 3D convolutions compared to 2D convolutions achieve better image quality [108]. However, one barrier is that such a method is computationally expensive. We utilize the spatial and temporal correlation in CT scans to introduce an efficient neural network architecture, Spatial-Temporal ResNet (STResNet) that restores the high-resolution details from low-dose CT images. Our goal is to achieve the same level of accuracy as the standard 3D SRResNet while improving its efficiency.

5.1.2 Method and data

Inspired by Enhanced Deep Residual Networks (EDSR) [73], we implemented a baseline denoising network based on SRResNet using fully 3D convolutional layers with a series of residual in residual blocks with convolutional and activation layers. Since CT scans are 3D volumes consisting of multiple slices, each slice can be treated as a frame at a time step. For each pixel in a slice, spatial and temporal correlation exists in adjacent frames along the temporal dimension. Hence, in STResNet, we decompose a full 3D convolution with $3 \times 3 \times 3$ kernel into two smaller convolutions, each with a spatial and temporal kernel. As illustrated in Figure 5.1, 3D convolutional blocks are replaced with spatial ($1 \times 3 \times 3$) and temporal

Figure 5.1: Convolutional blocks and results. A nodule ROI is highlighted in the circle.



$(3 \times 1 \times 1)$ convolutional blocks [68].

We demonstrate the differences in efficiency and accuracy using a dataset of low-dose CTs acquired for lung cancer screening, acquired at an equivalent dose about 2mGy. The standard condition was acquired at 100% dose and reconstructed using a medium kernel and 1.0 mm slice thickness, which reflects the parameters that are currently recommended for lung cancer screening. In the test set of 84 patients, 42 scans (50%) were found to have a total of 68 lung nodules. Lower-dose CT images were reconstructed from raw data of standard acquisitions using a physics-based model that simulates noise characteristics as well as reconstruction artifacts that are equivalent to 10% of the standard dose and at 2.0mm slice thickness. Data were split into 80/20/84 for training/validation/test. We adapted the NVIDIA APEX mixed-precision training package to further improve the training speed with mixed precision on GPU. We also introduced 8-bit low-precision quantization [50] to SRResNet and STResNet to achieve faster inference on CPU.

5.1.3 Evaluation and results

Our method was validated using image quality metrics such as peak signal-to-noise ratio (PSNR), structural similarity (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) [147]. In Table 5.1, our STResNet achieved better PSNR and SSIM compared to SRResNet in full precision (FP32) inference. Quantization (INT8) was shown to negatively impact image quality using PSNR and SSIM as metrics (a decrease of 0.16dB and 0.0152 respectively). However, compared to the baseline model, STResNet with 8-bit quantization achieved better perceptual quality (0.3555 vs. 0.3653). As shown in Figure 5.1, the difference between the result of baseline and 8-bit quantized STResNet at nodule ROI is visually imperceptible. During inference tasks on CPU, our quantized STResNet achieves up to 7.11 times speed-up compared to the standard SRResNet. Using STResNet alone achieves a speed up by a factor of 1.67. A similar trend is observed during training on GPU with up to 2 times speed up when using STResNet FP16 versus SRResNet FP32.

To assess differences in radiomic features, we selected 57 first-order intensity, gray-level concurrence matrix (GLCM), gray-level run length matrix (GLRLM), and gray-level size zone matrix (GLSZM) features to study the impact of feature values on nodules by using different combinations of networks and precision. In Figure 5.2, we found 54% of feature values from the outputs of quantized STResNet were still consistent to the baseline distribution.

Table 5.1: Image quality metrics and speed-up factors to baseline. * CPU results

		↑ PSNR(dB)	↑ SSIM	↓LPIPS	Inference time (sec)	Training time per iter (sec)	Inference Speed-up	Training Speed-up
FP32	SRResNet (baseline)	31.31±0.30	0.7216±0.0113	0.3635±0.0074	27.4(446.7*)	6.5	N/A	N/A
	STResNet	31.91±0.44	0.7265±0.0110	0.3715±0.0075	14.4(267.0*)	3.9	1.67	1.65
FP16	SRResNet	32.39±0.52	0.7277±0.0111	0.3640±0.0075	13.8	4.9	N/A	1.31
	STResNet	32.60±0.64	0.7259±0.0111	0.3732±0.0076	17.0	3.2	N/A	**2.04**
INT8	SRResNet	31.15±0.28	0.7064±0.0109	0.3501±0.0075	108.7*	N/A	4.11	N/A
	STResNet	31.11±0.30	0.7135±0.0109	0.3555±0.0076	62.8*	N/A	**7.11**	N/A

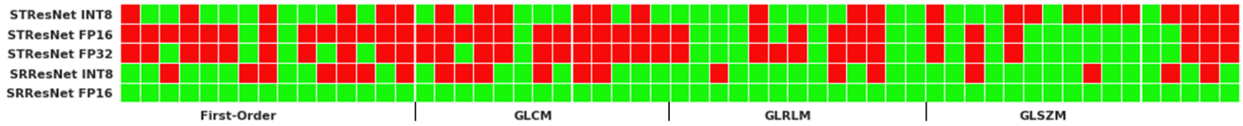


Figure 5.2: Radiomic features test. Red/Green indicates significant/non-significant difference to baseline via paired t-test with $p < 0.05$.

5.1.4 Discussion

We trained and evaluated our efficient and accurate network architecture called STResNet for low-dose CT denoising. Through our study, we demonstrated that STResNet reduces the training and inference time compared to SRResNet. We also showed that 8-bit quantization produced outputs that had minimal perceptual differences despite the information loss of computing a 12-bit CT scan using 8-bit quantized network weights. We note in our results that some radiomic features have statistically significant differences in distribution compared to feature values calculated from SRResNet outputs. Further study is required to assess the impact of 8-bit quantization and STResNet assumptions on downstream tasks such as machine learning algorithm performance. As part of future work, we will investigate the impact of using the efficient network architecture on clinical-driven tasks such as lung nodule detection or diffuse lung disease quantification.

5.2 Model Generalization

While quantitative image features (radiomics) can provide valuable information about disease progression, they are susceptible to variations in acquisition and reconstruction. Studies conducted previously have demonstrated that it is possible to normalize heterogeneous scans by using per-pixel metrics (e.g., mean squared error) and qualitative reader studies. Although these techniques are generalizable and may influence downstream tasks (e.g., classification), they have not been systematically studied. We present a multi-pronged evaluation by assessing image normalization techniques using 1) per-pixel image quality and perceptual metrics, 2) variability in radiomic features, and 3) task performance differences using a machine learning (ML) model. We evaluated the performance of a previously published 3D generative adversarial network (GAN) algorithm based on computed tomography (CT) scans acquired at different institutions with varying levels of radiation exposure. In spite of the superior metric results of the 3D GAN, its effects on quantitative image features and downstream performance were not universal. This study indicates a more complex relationship between CT acquisition and reconstruction parameters and their impact on radiomic features and ML model performance, which is not completely captured by per-pixel metrics alone. As a result of our analysis, we are able to provide a more comprehensive picture of the effect of normalization.

Radiomic features reflect small pixel- or voxel-wise changes that could be early indicators of disease progression but are not readily discernible by human readers [59]. However, these changes are often confounded by how images are acquired and reconstructed. Radiomic features are sensitive to scanner and acquisition parameters including dose, reconstruction kernel, and slice thickness [61,101]. Scanner-specific heterogeneity cannot be simply removed by the current preprocessing pipeline, as demonstrated in magnetic resonance imaging by Glocker et al. [35]. While efforts have examined ways to standardize acquisition, such solutions could only be applied prospectively and preclude the use of preexisting imaging data. Normalization techniques that reduce the variability due to CT acquisition and reconstruction would aid in the clinical translation of radiomic features.

Prior studies have examined how to normalize the heterogeneous scans computationally. In addition to traditional methods such as sinogram filtering [128] and iterative reconstruction [39], we highlight eight recent studies that employed neural networks to perform CT image enhancement or translation. Chen et al. [16] utilized a deep convolutional neural network (CNN) to transform low-dose CT (LDCT) images to appear as diagnostic dose. Wolterink et al. [135] coupled a CNN with an adversarial CNN to denoise the LDCT. Yi [141] proposed a generative adversarial network (GAN) that utilizes sharpness loss to leverage the blurry effect. Yang [140] demonstrated a GAN with Wasserstein (WGAN) and perceptual loss. You et al. [143] incorporated CNN with residual learning for superresolution image restoration. Aside from denoising, Liang et al. [71] and Selim et al. [107] proposed a GAN-based procedure to normalize CT images with different slice thicknesses and reconstruction kernels towards a predetermined standard. Our group proposed a 3D GAN model [132] that normalized dose and slice thickness simultaneously.

Table 5.2: Image quality metric results. A-G corresponds to a separate normalization scenario shown in Figure 5.3. D: dose level, K: kernel (sharp/smooth).

METRIC	**GAN**								**WGAN**
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	
	D: 10% → 100% K: smooth	D: 25% → 100% K: smooth	D: 10% → 100% K: sharp	D: 25% → 100% K: sharp	D: 10% → 100% K: smooth →sharp	D: 25% → 100% K: smooth →sharp	D: 10% → 100% K: sharp →smooth	D: 25% → 100% K: sharp →smooth	D: 10% → 100% K: sharp
PSNR	17.75	17.00	18.13	17.82	16.88	16.53	18.19	18.15	18.05
SSIM	0.6345	0.5360	0.6564	0.5914	0.4015	0.3460	0.6776	0.6788	0.6361
LPIPS (VGG)	0.2346	0.2445	0.1983	0.2057	0.3207	0.3481	0.2745	0.2574	0.2080

Out of these eight studies, six evaluated their approaches by computing image quality metrics such as peak signal-to-noise ratio (PSNR) and/or structural similarity (SSIM). Two employed subjective judgment by human readers. Two compared radiomic features. None examined the effect of normalization on downstream tasks. Six used training and test data from a single institution with the same type of scanners. Several limitations of these prior works are noted. First, PSNR and SSIM are computed on a per-pixel/voxel basis using low-level features, which fail to account for many nuances of perceptual similarity. Second,

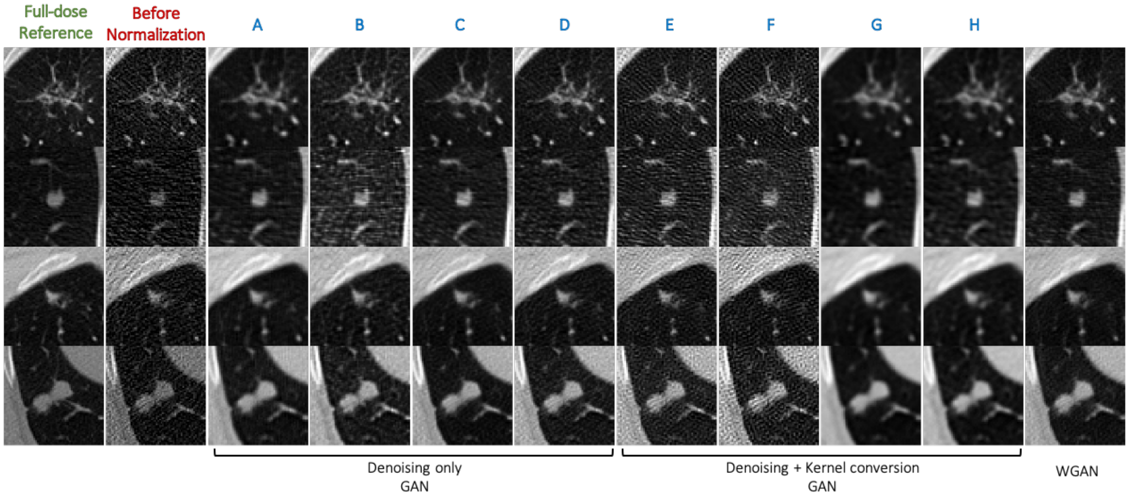


Figure 5.3: Visual comparison for four selected ROIs. Each row corresponds to one ROI that contains one annotated nodule. Each column corresponds to different normalization scenarios.

despite having the objective of improving detection, segmentation, or classification, studies have yet to demonstrate the effect of CT normalization on these tasks. Third, the generalizability of deep-learning-based normalization techniques has been understudied. Lack of external validation on scans acquired under different settings raises uncertainty over the actual model performance.

We propose a multi-pronged approach for evaluating normalization techniques to demonstrate the strengths and weaknesses of using image synthesis methods for CT normalization. Our approach's premise is that a single set of metrics does not provide a complete picture of the effect that normalization has on quantitative image features and downstream tasks. With a previously reported GAN-based normalization method, we assess the impact of normalization at three levels and the contributions of this work include 1) demonstrating the importance of image quality metrics beyond PSNR and SSIM; 2) highlighting the impact of normalization on radiomic features in critical anatomical regions of interest (ROIs); and 3) displaying the influence of normalization on the performance of a lung nodule detection algorithm.

5.2.1 Method

5.2.1.1 Dataset

25 chest LDCT scans from the Mayo Patient CT Projection Data Library [87] were used in this evaluation. Scans were acquired using a SOMATOM Definition Flash CT scanner (Siemens Healthineers, Erlangen, Germany) following standard clinical protocols. Simulated scans at 10% of the routine dose via inserting Poisson noise were available along with the original imaging data. The kernel used to reconstruct is B50f with a slice thickness of 1.0 mm. Radii (r) and centroids ([x, y, z]) of 42 annotated lesions were provided.

5.2.1.2 Model

We previously described a 3D GAN-based model for CT image normalization, whose architecture comprises a 3D SR-ResNet generator and a VGG-like discriminator with spectral normalization layers [132]. Eight models were trained independently based on different mappings (e.g., normalize from 10% dose, smooth kernel to 100% dose, sharp kernel), as specified in **Table 5.2**. A WGAN model based on [140] was also implemented for the purpose of comparison. These models were trained using images reconstructed from raw sinogram data for 80 patients extracted from Siemens CT scanners (Definition Flash, Sensation 64, Definition AS) at our institution. Based on a previously validated physics-based dose-reduction model [145], noise was injected into the raw data to simulate images with 10% and 25% of the acquisition dose. Final images were reconstructed using weighted filtered back projection [43] with a smooth or sharp kernel. The same training strategy was adopted from [132].

5.2.2 Overview of evaluation

The evaluation was conducted in three parts: 1) image quality assessment using per-pixel metrics (PSNR, SSIM) and a high-level perceptual similarity metric; 2) radiomic feature

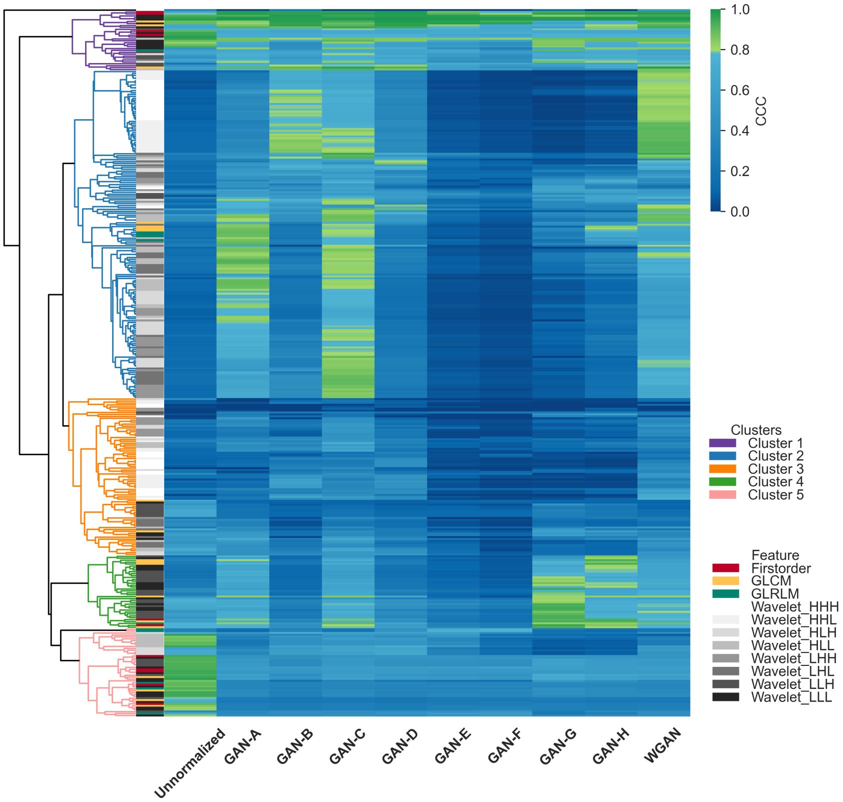


Figure 5.4: Heatmap and dendrogram displaying the level of agreement among radiomic features. A CCC ≥ 0.8 was considered as indication for high agreement (shaded green). Features were loosely clustered into five groups using average linkage algorithm. H = high-pass filtering. L = low-pass filtering. Filters were applied successively in x, y, z direction (e.g., "HLH").

analysis on absolute feature errors and reproducibility; and 3) task-dependent assessment using a pretrained nodule detection algorithm. We performed these evaluations by applying 3D GAN-based models trained at our institution to unseen scans from the Mayo Clinic dataset.

5.2.2.1 Image quality assessment

PSNR and SSIM assess global per-pixel/voxel differences but can easily fail to reflect differences which are apparent to human vision. We thus employed another metric named

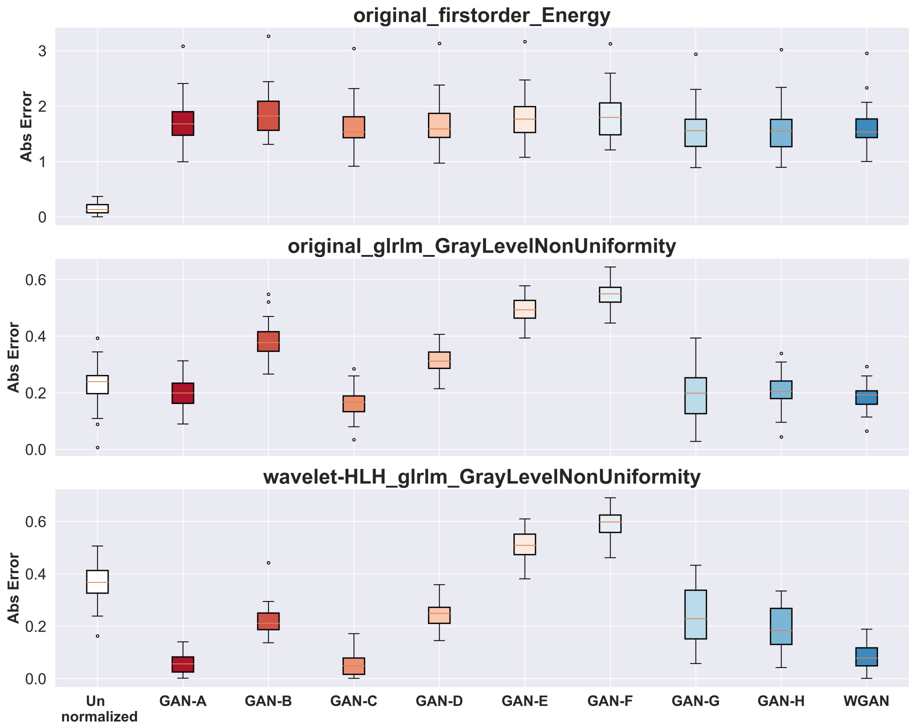


Figure 5.5: Absolute errors in three representative radiomic features

Learned Perceptual Image Patch Similarity (LPIPS) [146], which evaluates semantic similarity using deep features computed by a pretrained VGG-16 model (version=0.0). The network was trained on a perceptual similarity dataset (Berkeley-Adobe Perceptual Patch Similarity Dataset [146]) which contains 484k human perceptual judgments. A lower LPIPS value reflects a closer perceptual distance. We pursued a task-based approach by focusing on ROIs with anatomical significance (e.g., lung parenchyma). Calculated along the axial plane, metric results were averaged over 42 annotated ROIs with a dimension of 64 × 64 for each scenario.

5.2.2.2 Radiomic feature analysis

Motivated by the quantitative radiomics analysis presented in [5], a total of 369 radiomic features were extracted from 42 ROIs of dimension 64 × 64 × 5 with pyradiomics [126], which includes: (i) 11 intensity features based on first-order statistics, (i) 30 texture features describing spatial distribution of voxel intensities and (iii) aforementioned 41 features from eight wavelet decompositions of original images by directional (x, y, z) low-pass and high-pass filtering. We computed absolute errors for each radiomic feature by $|\hat{\mathbf{x}}_i - \mathbf{x}_i|/\mathbf{x}_i$ where $\hat{\mathbf{x}}_i$ refers to feature value without/after normalization and $\mathbf{x}_i$ is the feature value in reference. Lin's concordance correlation coefficient (CCC) [62] was employed to study reproducibility of radiomic features after normalization.

5.2.2.3 Nodule detection performance

We employed a 2D RetinaNet trained on the LIDC-IDRI dataset [74] to perform nodule detection on the 25 scans containing all annotated lesions. Lobe segmentation by U-Net [46] was applied to minimize interference from outer chest bone structures. With the provided centroid and radius (r), we generated a spherical ROI for each nodule. Slices whose intersection with the ROI has a radius greater than 0.95r were also adopted as the reference standard. The model returns a set of rectangular bounding boxes, each with a probability of containing a nodule. A bounding box with a probability > 0.5 is considered as a nodule and as a true positive if its intersection over union with ground truth exceeds 0.5. Since a detection model may have its own inherent performance limitations, we computed CCCs for precision and recall between each normalization scenario and the reference.

5.2.3 Results

5.2.3.1 Image quality metrics

Results for image quality metrics with a visual comparison of four ROIs were presented in **Table 5.2** and **Figure 5.3**. Among four denoising-only scenarios, **A** and **C** with 10% to 100% dose mappings, respectively, achieved a higher score than **B** and **D** with 25% to 100% mapping. Better PSNR and LPIPS were observed for models trained on images reconstructed with sharp kernel (**C** and **D**). All four scenarios involving kernel conversion resulted in poorer LPIPS. A noticeable drop in all three metrics was observed when attempting to convert a smooth kernel to a sharp kernel, whose ROIs in **Figure 5.3** displayed sharpening of edges with a high noise level. In comparison, **G** and **H** achieved comparable PSNR and SSIM results with the denoising-only scenarios but failed in LPIPS. Their ROIs also visually appeared oversmoothed.

5.2.3.2 Radiomic features

Figure 5.4 provides an overview of reproducibility of radiomic features after normalization. Among the five clusters, an increase in agreement was observed for the majority of features in Cluster 2 for **A**-**D**. Features in Cluster 4 also displayed general improvement in **A** and **C**, but also **G** and **H**. In contrast, features in Cluster 3 and 5 underwent unimproved or worsened agreement under all normalization scenarios. Features in Cluster 1 were characterized by consistently high concordance with the reference throughout the normalization process. **Figure 5.5** shows the absolute errors in three radiomic features with high prognostic power for lung cancer [5]. Similar trends for different scenarios were observed except the increase in absolute errors for the first-order feature "Energy" extracted from original images.

5.2.3.3 Nodule detection performance

Figure 5.6 summarizes the relative performance of nodule detection between different normalization scenarios and a 100% dose reference. Agreement for precision was generally improved after normalization except for **E** and **F** while significant improvement was only observed in **A**, **C** and **G** for recall.

5.2.4 Discussion

The inadequacy of conventional metrics (PSNR, SSIM) in assessing visual similarity implies the potential benefits for employing diverse metrics during performance evaluation. Although the images in **G** and **H** appeared oversmoothed, their PSNR and SSIM were still comparable with images more perceptually close to the reference. Such discrepancy was captured using

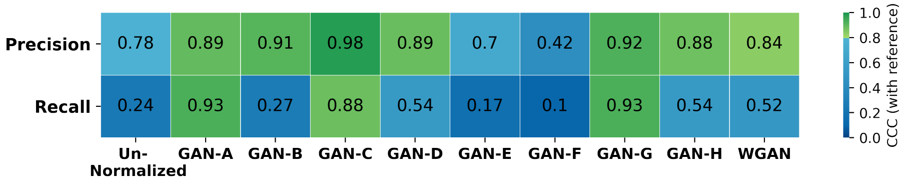


Figure 5.6: Nodule detection performance. We applied CCC to compare the relative performance of different normalization scenarios (precision & recall) with respect to the full-dose reference.

LPIPS, one option among many high-level metrics [33, 98]. Nevertheless, none of these metrics alone was adequate to predict the effects on radiomic features and ML model outputs as presented. Results from different metrics and direct visual comparison are thus essential for understanding the impact of normalization techniques.

From the metric results and direct visualization, we observed that certain types of mappings in CT normalization are more challenging than others. Kernel conversion remains critical especially when generalized on external datasets. Conversion using our 3D GAN from smooth to sharp kernel appears to add ringing artifacts into the images (**E** and **F** in **Figure 5.3**). Whether such difficulties originate from specific neural network characteristics and potential mitigation strategies demand further investigation.

While overall noise reduction is achieved, the normalized intensity and textural values highlighted by radiomic features are not necessarily equivalent to those generated from reference. As seen in **Figure 5.4**, a fair proportion of features cannot be easily improved through normalization and are highly sensitive to the changes in reconstruction kernels. Using a single radiomic signature could also result in an incomplete picture about the effects of normalization, as shown in **Figure 5.5**.

Although image quality and radiomic features appear as strong contributors to improvement in nodule detection performance on normalized images, this process is not deterministic and might depend on how different radiomic features are combined and weighted. As images normalized by the **WGAN** model achieved equivalently good metric and radiomic results as **GAN-C**, it did not result in as much improvement in both precision and recall. Rather than treating as separate processes, optimizing the overall performance for CT normalization and downstream tasks (detection, segmentation, classification) could potentially be more efficient.

In summary, we demonstrate that a single set of image quality metrics is insufficient to predict whether a normalization technique has beneficial impacts on radiomics and downstream tasks. Adopting a multi-pronged approach provides a more complete understanding of the effect of normalization, particularly related to the use of these scans in clinical tasks. For future work, we wish to translate this idea into more quantitative metrics for performance comparison between different normalization techniques.

CHAPTER 6

Conclusion

6.1 Summary

The purpose of this dissertation is to develop and compare two modeling approaches for CT image normalization, a conditional GAN approach presented in Chapter 3 and a conditional Flow approach presented in Chapter 4. We examine quantitative imaging features and task-driven clinical evaluations beyond the use of image quality metrics in Chapter 5. Few works in this field are presented with a holistic evaluation. The use of a single set of image quality metrics is insufficient to show whether a normalization technique benefits downstream tasks. A multi-pronged approach provides a deeper understanding of how normalization impacts downstream tasks and how it may support clinical decisions.

There are limitations to GAN-based approaches. The GAN may introduce undesired artifacts in the radiomic feature analysis, where unnormalized images may, in some circumstances, perform better than normalized images. GANs also models the normalization problem deterministically, even though the problem is intrinsically ill-posed, and multiple outputs could be valid. This limitation was partially addressed in Chapter 4 with the Flow-based method, which illustrates the benefits of a flow-based model over a popular GAN-based model and the impact on ensuring consistent performance of CAD software. A unique feature of the flow-based model is its ability to measure uncertainty by using different output images that are obtained by sampling the latent space. CTFlow delivers more reliable lung nodule detection results by using explicit density. We can leverage latent space autoencoding to make models generalizable. The model uncertainty estimation offered by normalizing flows can also be extended to other medical image analysis beyond prepossessing images, such as segmentation, detection, and classification. However, CTFlow is not universally superior to SNGAN. In Chapter 4.4.2, we showed that SNGAN outperforms direct CTFlow for the 10% dose evaluation in Mayo datasets. When dealing with low-dose denoising problems, SNGAN demonstrates its strength in generating samples when the conditioned image is of inferior quality.

Moreover, certain input conditions and feature combinations are robust to variations in CT parameters. Therefore, applying normalization techniques is unnecessary and may even amplify artifacts. Some quantitative imaging features are robust under a variety of conditions, but there may not be a condition under which all features are robust.

Studies in the past have examined the effect of individual CT parameters (e.g., dose, slice thickness, kernel) on feature reproducibility but have neglected to take into account the interactions among CT acquisition and reconstruction parameters. For example, if the effect of radiation dose reduction is examined at a specific kernel and slice thickness, we may overlook their additive impacts on noise. Therefore, in the future, we will seek to understand the complex nature of these effects on different clinical tasks systematically and comprehensively and conduct a breakdown analysis of those features and conditions into classifications requiring aggressive, moderate, or no mitigation.

The term “artifact” refers to patterns, textures, features, and morphological structures that are introduced into the normalized image but are not tied to physical reality. As can be seen from GAN-based normalization results, artifacts are introduced into the reconstructed image. Normalization based on Flow is effective when dealing with artifacts. Nonetheless, it is necessary to investigate whether these artifacts originate from the model or the training set. The distribution of samples with abnormalities (lesions) in the training set must match the distribution of abnormalities in the test set. In other words, a good model must be well calibrated. It is recommended that the probability of generating a lesion or abnormality matches the actual probability of seeing lesions in the test set, to ensure model reliability. The flow-based approach has the potential to provide probability information so that it can be calibrated manually in accordance with the true probability observed in the test set.

We expect to see more work estimating uncertainty and utilizing it to improve classification, segmentation, and detection for medical image normalization. Flow-based approaches facilitate the interpretation of models by providing clinicians with a sense of model confidence. However, GANs will continue to dominate image generation tasks. Meanwhile, the topic of overcoming mode collapse is an ongoing research to increase sample diversity for GAN. When should one consider using GAN over Flow? To answer this question, we must first consider the fundamental differences between the two approaches: GAN models implicit density, whereas Flow learns explicit density. While GANs have the advantage of generating visually appealing outputs, Flow learns the optimal likelihood. As discussed in Chapter 4.25, explicit density can be used to manipulate images. GANs can provide useful outputs, if having an artifact-free image is not a concern (e.g., medical image synthesis for data augmentation). Flow might be a better choice for medical images, but GAN is still superior for natural images.

6.2 Concluding remarks

There are open challenges in the normalization of medical images. At the dataset level, all models are trained with paired data in this research. Paired data can either be acquired using ground truth imaging data or by simulating lower-dose or nonstandard imaging conditions. In the context of image normalization, to obtain a “gold standard” ground truth training dataset, we would ideally need to acquire paired normal dose and lower dose scans on the same patient. While such acquisitions do occur in very specific circumstances, requiring patients to be exposed to radiation unnecessarily would be unethical. A simulated approach such as the one that we pursued has several limitations: 1) The intra- and inter-variability consists of multiple parameter variations resulting in combinations of imaging conditions; accordingly, it is impossible to enumerate all these conditions through simulation. 2) Simulated scenarios and real-world situations are always different. The challenge is to develop a physical model that captures every physical process of the image variation due to different CT parameters. Though in this study, the simulated dataset was based on a sophisticated physics-based model that had previously been compared and validated statistically with the characteristics of the ground truth phantom scan, in most cases, simulated data do not exhibit the same characteristics. However, paired data may not even be necessary. Recent advances, such as CycleGAN, do not require paired data. This allows researchers to scale up training with more data by utilizing a large volume of unpaired data. Additionally, this idea can be extended by using more training data to reduce variability as well as directly training the downstream image analysis model using imaging data with mixed conditions. As a result, we refer to this approach as data augmentation, which makes the model more generalizable. This method is intended to use more diverse conditions or parameter combinations within the dataset to act as a regularizer and reduce overfitting to limited imaging conditions.

Ultimately, the normalization and the data augmentation approach are similar. This dissertation investigated the effect of different normalization techniques on mitigating the effects of different CT conditions on quantitative imaging features. Under the current settings of imaging data accessibility, it is difficult to pursue data augmentation due to two reasons: 1) AI/ML developers and healthcare institutions are unable to retrain their models on a timely basis due to limited data availability and proprietary software and 2) the downstream image analysis model differs based on the specific task and type of disease. Retraining a model for every possible environment that the model could be used is neither feasible nor scalable.

One additional point: despite the excitement surrounding the potential use of AI/ML for medical image analysis, clinical practice has not fundamentally changed. I believe one major impediment is the lack of good quality training data. In the healthcare industry, legislation and policies regarding data privacy and security represent a growing concern, and therefore data are limited.

However, there are two approaches to overcome privacy concerns. With federated learning, researchers can train statistical models on decentralized devices or servers utilizing local data sets. Researchers can train using the same model without uploading private data to the cloud or exchanging data with other parties. By maintaining local data storage, federated learning reduces data security and privacy risks relative to traditional machine learning operating in a centralized manner. The parameters of the model are later merged into a central node. However, one central node responsible for orchestrating all models is prone to security breaches. The second approach, swarm learning, is also a decentralized and confidential clinical machine learning scheme introduced in [130] that sheds some light on how ML distributed computing would eventually evolve to for healthcare industry. Peer-to-peer networking and coordination are performed through blockchain technology, while confidentiality is maintained without a central node. This ensures the security of the network through proof-of-work [26]. As a scientific demonstration, Warnat-Herresthal et. al. shows that Swarm Learning classifiers outperform those developed at individual sites in predicting radiological findings (atelectasis, effusion, infiltration) for a public dataset containing 95,000 chest X-ray images. The goals of this dissertation could have been also achieved through decentralized federated or swarm learning. For image variability caused by multicenter effect, each center can train the model using its own data, but merge network parameters later using federated or swarming learning. The model will be more robust against the effects of variations in data acquisition protocols with the help of larger datasets from other centers.

The future of healthcare is to employ the aforementioned technologies to enable the sharing of deidentified medical data amongst institutions, to build robust multicenter data sharing platforms, and ultimately to train large-scale models using massive medical datasets. These efforts are vital to the creation of better tools and infrastructures for model development. We will continue to face challenges in developing models from limited data until the healthcare infrastructure matures in the future.

REFERENCES

[1] LUNG CANCER SCREENING CT. (American Association of Physicists in Medicine, 2019).

[2] Lung cancer screening ct protocols version, 2019.

[3] AAPM. “low dose ct grand challenge. *Online*, 2017.

[4] Jonas Adler and Ozan Öktem. Deep posterior sampling: Uncertainty quantification for large scale inverse problems. In *International Conference on Medical Imaging with Deep Learning–Extended Abstract Track*, 2019.

[5] H. Aerts, E. Velazquez, and R. Leijenaar et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communication*, Jun 2014.

[6] Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, 5(1):1–9, 2014.

[7] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322, 2006.

[8] Diego Ardila, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, 25(6):954–961, 2019.

[9] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

[10] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, and Laurence P Clarke. Data from lidc-idri. *The cancer imaging archive*, 10:K9, 2015.

[11] Joanne C. Beer, Nicholas J. Tustison, Philip A. Cook, Christos Davatzikos, Yvette I. Sheline, Russell T. Shinohara, and Kristin A. Linn. Longitudinal combat: A method for harmonizing longitudinal multi-scanner imaging data. *NeuroImage*, 220:117129, 2020.

[12] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Meskó. The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database. *NPJ digital medicine*, 3(1):1–8, 2020.

[13] F Edward Boas, Dominik Fleischmann, et al. Ct artifacts: causes and reduction techniques. *Imaging Med*, 4(2):229–240, 2012.

[14] Matthew S Brown, Pechin Lo, Jonathan G Goldin, Eran Barnoy, Grace Hyun J Kim, Michael F McNitt-Gray, and Denise R Aberle. Toward clinically usable cad for lung cancer screening with computed tomography. *European radiology*, 24(11):2719–2728, 2014.

[15] Alex Chan, Ahmed Alaa, Zhaozhi Qian, and Mihaela Van Der Schaar. Unlabelled data improves bayesian uncertainty calibration under covariate shift. In *International Conference on Machine Learning*, pages 1392–1402. PMLR, 2020.

[16] H. Chen, Y. Zhang, M.K. Kalra, F. Lin, Y. Chen, P. Liao, J. Zhou, and G. Wang. Low dose ct with a residual encoder-decoder convolutional neural network. *IEEE Transactions on Medical Imaging*, 36(12):2524–2535, Jun 2017.

[17] Hu Chen, Yi Zhang, Mannudeep K Kalra, Feng Lin, Yang Chen, Peixi Liao, Jiliu Zhou, and Ge Wang. Low-dose ct with a residual encoder-decoder convolutional neural network. *IEEE transactions on medical imaging*, 36(12):2524–2535, 2017.

[18] Yuhua Chen, Feng Shi, Anthony G Christodoulou, Yibin Xie, Zhengwei Zhou, and Debiao Li. Efficient and accurate mri super-resolution using a generative adversarial network and 3d multi-level densely connected network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 91–99. Springer, 2018.

[19] Jooae Choe, Sang Min Lee, Kyung-Hyun Do, Gaeun Lee, June-Goo Lee, Sang Min Lee, and Joon Beom Seo. Deep learning–based image conversion of ct reconstruction kernels improves radiomics reproducibility for pulmonary nodules or masses. *Radiology*, 292(2):365–373, 2019.

[20] Jae Kyu Choi, Bin Dong, and Xiaoqun Zhang. Limited tomography reconstruction via tight frame and simultaneous sinogram extrapolation. *Journal of Computational Mathematics*, 34(6):575, 2016.

[21] Gary JR Cook, Connie Yip, Muhammad Siddique, Vicky Goh, Sugama Chicklore, Arunabha Roy, Paul Marsden, Shahreen Ahmad, and David Landau. Are pretreatment 18f-fdg pet tumor textural features in non–small cell lung cancer associated with response and survival after chemoradiotherapy? *Journal of nuclear medicine*, 54(1):19–26, 2013.

[22] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007.

[23] Alexander Denker, Maximilian Schmidt, Johannes Leuschner, Peter Maass, and Jens Behrmann. Conditional normalizing flows for low-dose computed tomography image reconstruction. *ICML Workshop on Invertible Neural Networks*, 2020.

[24] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: non-linear independent components estimation. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

[25] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[26] Cynthia Dwork and Moni Naor. Pricing via processing or combatting junk mail. In *Annual international cryptology conference*, pages 139–147. Springer, 1992.

[27] Idris A Elbakri and Jeffrey A Fessler. Statistical image reconstruction for polyenergetic x-ray computed tomography. *IEEE transactions on medical imaging*, 21(2):89–99, 2002.

[28] Nastaran Emaminejad, Pechin Lo, Shahnaz Ghahremani, Grace H Kim, Matthew S Brown, and Michael F McNitt-Gray. The effects of slice thickness and radiation dose level variations on computer-aided diagnosis (cad) nodule detection performance in pediatric chest ct scans. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, page 101340B. International Society for Optics and Photonics, 2017.

[29] Nastaran Emaminejad, Muhammad Wahi-Anwar, John Hoffman, Grace H. Kim, Matthew S. Brown, and Michael McNitt-Gray. The effects of variations in parameters and algorithm choices on calculated radiomics feature values: initial investigations and comparisons to feature variability across CT image acquisition conditions. In Nicholas Petrick and Kensaku Mori, editors, *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, pages 886 – 895. International Society for Optics and Photonics, SPIE, 2018.

[30] Nastaran Emaminejad, Muhammad Wasil Wahi-Anwar, Grace Hyun J Kim, William Hsu, Matthew Brown, and Michael McNitt-Gray. Reproducibility of lung nodule radiomic features: Multivariable and univariable investigations that account for interactions between ct acquisition and reconstruction parameters. *Medical Physics*, 2021.

[31] P Fumene Feruglio, Claudio Vinegoni, J Gros, A Sbarbati, and R Weissleder. Block matching 3d random noise filtering for absorption optical projection tomography. *Physics in Medicine & Biology*, 55(18):5401, 2010.

[32] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[33] F. Gao, Y. Wang, P. Li, M. Tan, J. Yu, and Y. Zhu. Deepsim: Deep similarity for image quality assessment. *Neurocomputing*, 257:104 – 114, 2017.

[34] Lucas L Geyer, U Joseph Schoepf, Felix G Meinel, John W Nance Jr, Gorka Bastarrika, Jonathon A Leipsic, Narinder S Paul, Marco Rengo, Andrea Laghi, and Carlo N De Cecco. State of the art: iterative ct reconstruction techniques. *Radiology*, 276(2):339–357, 2015.

[35] B. Glocker, R. Robinson, D.C. Castro, Q. Dou, and E. Konukoglu. Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects. In *NeurIPS Workshop*, Vancouver, Canada, Dec 2019.

[36] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[37] Aditya Grover, Manik Dhar, and Stefano Ermon. Flow-gan: Combining maximum likelihood and adversarial learning in generative models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[38] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[39] A.K. Hara, R.G. Paden, A.C. Silva, J.L. Kujak, H.J. Lawder, and W. Pavlicek. Iterative reconstruction technique for reducing body radiation dose at ct: feasibility study. *American Journal of Roentgenology*, 193(3):764–71, Sep 2009.

[40] Andreas Hauptmann and Ben T Cox. Deep learning in photoacoustic tomography: Current approaches and future directions. *Journal of Biomedical Optics*, 25(11):112903, 2020.

[41] Brendan F. Hayden. Slice reconstruction.

[42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[43] J. Hoffman, S. Young, F. Noo, and M. McNitt-Gray. Technical note: Freect_wfbp: A robust, efficient, open-source implementation of weighted filtered backprojection for helical, fan-beam ct. *Medical Physics*, 43(3), 3 2016.

[44] John Hoffman, Nastaran Emaminejad, Muhammad Wahi-Anwar, Grace H Kim, Matthew Brown, Stefano Young, and Michael McNitt-Gray. Design and implementation of a high-throughput pipeline for reconstruction and quantitative analysis of ct image data. *Medical physics*, 46(5):2310–2322, 2019.

[45] John Hoffman, Stefano Young, Frédéric Noo, and Michael McNitt-Gray. Freect_wfbp: A robust, efficient, open-source implementation of weighted filtered backprojection for helical, fan-beam ct. *Medical physics*, 43(3):1411–1420, 2016.

[46] J. Hofmanninger, F. Prayer, J. Pan, S. Rohrich, H. Proschm, and G. Langs. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European Radiology Experimental*, 4(50), Aug 2020.

[47] Clément Hognon, Florent Tixier, Olivier Gallinato, Thierry Colin, Dimitris Visvikis, and Vincent Jaouen. Standardization of multicentric image datasets with generative adversarial networks. In *IEEE Nuclear Science Symposium and Medical Imaging Conference 2019*, 2019.

[48] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[49] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[50] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2704–2713, 2018.

[51] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 04 2006.

[52] Daniel E Jonas, Daniel S Reuland, Shivani M Reddy, Max Nagle, Stephen D Clark, Rachel Palmieri Weber, Chineme Enyioha, Teri L Malo, Alison T Brenner, Charli Armstrong, et al. Screening for lung cancer with low-dose computed tomography: updated evidence report and systematic review for the us preventive services task force. *Jama*, 325(10):971–987, 2021.

[53] Jayashree Kalpathy-Cramer, Artem Mamomov, Binsheng Zhao, Lin Lu, Dmitry Cherezov, Sandy Napel, Sebastian Echegaray, Daniel Rubin, Michael McNitt-Gray, Pechin Lo, et al. Radiomics of lung nodules: a multi-institutional study of robustness and agreement of quantitative imaging features. *Tomography*, 2(4):430, 2016.

[54] Dongwoo Kang, Piotr Slomka, Ryo Nakazato, Jonghye Woo, Daniel S Berman, C-C Jay Kuo, and Damini Dey. Image denoising of low-radiation dose coronary ct angiography by an adaptive block-matching 3d algorithm. In *Medical Imaging 2013: Image Processing*, volume 8669, page 86692G. International Society for Optics and Photonics, 2013.

[55] A Karahaliou, S Skiadopoulos, I Boniatis, P Sakellaropoulos, E Likaki, G Panayiotakis, and L Costaridou. Texture analysis of tissue surrounding microcalcifications on mammograms for breast cancer diagnosis. *The British journal of radiology*, 80(956):648–656, 2007.

[56] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[57] Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[58] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[59] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R.G. Van Stiphout, P. Granton, C.M.L. Zegers, R. Gillies, R. Boellard, A. Dekker, and H.J. Aerts. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer*, 48(4):441–446, Mar 2012.

[60] Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud GPM Van Stiphout, Patrick Granton, Catharina ML Zegers, Robert Gillies, Ronald Boellard, André Dekker, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer*, 48(4):441–446, 2012.

[61] R. Larue, J.E. van Timmeren, E. de Jong, G. Feliciani, R. Leije-naar, W. Schreurs, M.N. Sosef, F. Raat, F. van der Zande, M. Das, W. van Elmpt, and P. Lambin. Influence of gray level discretization on radiomic feature stability for different ct scanners, tube currents and slice thickness-es: a comprehensive phantom study. *Acta Oncologica*, 57(11):1475–1481, Sep 2017.

[62] I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.

[63] A Lecler, L Duron, D Balvay, J Savatovsky, O Bergès, M Zmuda, E Farah, O Galatoire, A Bouchouicha, and LS Fournier. Combining multiple magnetic resonance imaging sequences provides independent reproducible radiomics features. *Scientific reports*, 9(1):1–8, 2019.

[64] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[65] Sang Min Lee, June-Goo Lee, Gaeun Lee, Jooae Choe, Kyung-Hyun Do, Namkug Kim, and Joon Beom Seo. Ct image conversion among different reconstruction kernels without a sinogram by using a convolutional neural network. *Korean journal of radiology*, 20(2):295–303, 2019.

[66] Robert M Lewitt. Multidimensional digital image representations using generalized kaiser–bessel window functions. *JOSA A*, 7(10):1834–1846, 1990.

[67] Meng Li, William Hsu, Xiaodong Xie, Jason Cong, and Wen Gao. Sacnn: Self-attention convolutional neural network for low-dose ct denoising with self-supervised perceptual loss network. *IEEE transactions on medical imaging*, 39(7):2289–2301, 2020.

[68] Sheng Li, Fengxiang He, Bo Du, Lefei Zhang, Yonghao Xu, and Dacheng Tao. Fast spatio-temporal residual network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10522–10531, 2019.

[69] Tianqing Li, Leihao Wei, and William Hsu. A multi-pronged evaluation for image normalization techniques. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1292–1296. IEEE, 2021.

[70] Zhoubo Li, Lifeng Yu, Joshua D Trzasko, David S Lake, Daniel J Blezek, Joel G Fletcher, Cynthia H McCollough, and Armando Manduca. Adaptive nonlocal means filtering based on local noise level for ct denoising. *Medical physics*, 41(1):011908, 2014.

[71] G. Liang, S. Fouladvand, J. Zhang, M. A. Brooks, N. Jacobs, and J. Chen. Ganai: Standardizing ct images using generative adversarial network with alternative improvement. In *IEEE International Conference on Healthcare Informatics*, pages 1–11, Nov 2019.

[72] Gongbo Liang, Sajjad Fouladvand, Jie Zhang, Michael A Brooks, Nathan Jacobs, and Jin Chen. Ganai: standardizing ct images using generative adversarial network with alternative improvement. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–11. IEEE, 2019.

[73] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[74] Y. Lin, L. Wei, S.X. Han, D.R. Aberle, and W. Hsu. Edicnet: an end-to-end detection and interpretable classification network for pulmonary nodules on computed tomography. *Proceedings of SPIE–the International Society for Optical Engineering*, 11314, Feb 2020.

[75] Yannan Lin, Leihao Wei, Simon X Han, Denise R Aberle, and William Hsu. Edicnet: An end-to-end detection and interpretable malignancy classification network for pulmonary nodules in computed tomography. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, page 113141H. International Society for Optics and Photonics, 2020.

[76] P Lo, S Young, HJ Kim, MS Brown, and MF McNitt-Gray. Variability in ct lung-nodule quantification: effects of dose reduction and reconstruction methods on density and texture based features. *Medical physics*, 43(8Part1):4854–4865, 2016.

[77] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[78] Lin Lu, Ross C Ehmke, Lawrence H Schwartz, and Binsheng Zhao. Assessing agreement between radiomic features computed for multiple ct imaging settings. *PloS one*, 11(12):e0166550, 2016.

[79] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Srflow: Learning the super-resolution space with normalizing flow. In *European Conference on Computer Vision*, pages 715–732. Springer, 2020.

[80] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[81] R. Mahon, M. Ghita, Geoffrey D. Hugo, and E. Weiss. Combat harmonization for radiomic features in independent phantom and lung cancer patient computed tomography datasets. *Physics in medicine and biology*, 2019.

[82] Liting Mao, Huan Chen, Mingzhu Liang, Kunwei Li, Jiebing Gao, Peixin Qin, Xianglian Ding, Xin Li, and Xueguo Liu. Quantitative radiomic model for predicting malignancy of small solid pulmonary nodules detected by low-dose ct screening. *Quantitative imaging in medicine and surgery*, 9(2):263, 2019.

[83] Maria D Martin, Jeffrey P Kanne, Lynn S Broderick, Ella A Kazerooni, and Cristopher A Meyer. Lung-rads: pushing the limits. *Radiographics*, 37(7):1975–1993, 2017.

[84] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[85] M. Mazurowski. Radiogenomics: what it is and why it is important. *Journal of the American College of Radiology : JACR*, 12 8:862–6, 2015.

[86] G.B. McBride. A proposal for strength-of-agreement criteria for lin's concordance correlation coefficient. *NIWA Client Report: HAM2005-062.*, 2005.

[87] C.H. McCollough, B. Chen, D. III Holmes, X. Duan, Z. Yu, L. Yu, S. Leng, and J. Fletcher. Low dose ct image and projection data [data set]. *The Cancer Imaging*, 2020.

[88] Paulius Micikevicius. Mixed-precision training of deep neural networks. *NVidia White Paper*, 2017.

[89] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[90] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[91] Wei Mu, Ilke Tunali, Jhanelle E Gray, Jin Qi, Matthew B Schabath, and Robert J Gillies. Radiomics of 18 f-fdg pet/ct images predicts clinical benefit of advanced nsclc patients to checkpoint blockade immunotherapy. *European journal of nuclear medicine and molecular imaging*, 47(5):1168–1182, 2020.

[92] Fanny Orlhac, Sarah Boughdad, C. Philippe, Hugo Stalla-Bourdillon, C. Nioche, L. Champion, M. Soussan, F. Frouin, V. Frouin, and I. Buvat. A postreconstruction harmonization method for multicenter radiomic studies in pet. *The Journal of Nuclear Medicine*, 59:1321 – 1328, 2018.

[93] Vishwa Parekh and Michael A Jacobs. Radiomics: a new application from established techniques. *Expert review of precision medicine and drug development*, 1(2):207–226, 2016.

[94] Junyoung Park, Donghwi Hwang, Kyeong Yun Kim, Seung Kwan Kang, Yu Kyeong Kim, and Jae Sung Lee. Computed tomography super-resolution using deep convolutional neural network. *Physics in Medicine & Biology*, 63(14):145011, 2018.

[95] Chintan Parmar, Ralph TH Leijenaar, Patrick Grossmann, Emmanuel Rios Velazquez, Johan Bussink, Derek Rietveld, Michelle M Rietbergen, Benjamin Haibe-Kains, Philippe Lambin, and Hugo JWL Aerts. Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. *Scientific reports*, 5(1):1–10, 2015.

[96] Nick Pawlowski and Ben Glocker. Abnormality detection in histopathology via density estimation with normalising flows. In *International Conference on Medical Imaging with Deep Learning–Short Papers Track*, 2021.

[97] K. Pinker, F. Shitano, E. Sala, R. Do, R. Young, A. Wibmer, H. Hricak, E. Sutton, and E. Morris. Background, current role, and potential applications of radiogenomics. *Journal of Magnetic Resonance Imaging*, 47, 2018.

[98] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *CVPR*, 2018.

[99] Harvard Health Publishing. Radiation risk from medical imaging - harvard health, 04 2018.

[100] Sathish Ramani and Jeffrey A Fessler. A splitting-based iterative algorithm for accelerated statistical x-ray ct reconstruction. *IEEE transactions on medical imaging*, 31(3):677–688, 2011.

[101] S. Rizzo, F. Botta, S. Raimondi, D. Origgi, C. Fanciullo, A.G. Morganti, and M. Bellomi. Radiomics: the facts and the challenges of image analysis. *European radiology experimental*, 2(1):36, Nov 2018.

[102] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[103] Da-ano Ronrick, Ingrid Masson, Lucia Francois, Caroline Rousseau, Augustin Mervoyer, Pietro Bonaffini, Caroline Reinhold, Selima Sellami, Joel Castelli, Renaud De Crevoisier, et al. Performance comparison of modified combat for harmonization of radiomic features in multicentric studies, 2020.

[104] Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, Christian Wachinger, Alzheimer's Disease Neuroimaging Initiative, et al. Bayesian quicknat: model uncertainty in deep whole-brain segmentation for structure-wise quality control. *NeuroImage*, 195:11–22, 2019.

[105] Romke Rozema, Herbert T Kruitbosch, Baucke van Minnen, Bart Dorgelo, Joep Kraeima, and Peter MA van Ooijen. Iterative reconstruction and deep learning algorithms for enabling low-dose computed tomography in midfacial trauma. *Oral surgery, oral medicine, oral pathology and oral radiology*, 132(2):247–254, 2021.

[106] Jo Schlemper, Daniel C Castro, Wenjia Bai, Chen Qin, Ozan Oktay, Jinming Duan, Anthony N Price, Jo Hajnal, and Daniel Rueckert. Bayesian deep learning for accelerated mr image reconstruction. In *International Workshop on Machine Learning for Medical Image Reconstruction*, pages 64–71. Springer, 2018.

[107] M. Selim, J. Zhang, B. Fei, G.Q. Zhang, and J. Chen. Stan-ct: Standardizing ct image using generative adversarial network. *arXiv*, Apr 2020.

[108] Hongming Shan, Yi Zhang, Qingsong Yang, Uwe Kruger, Mannudeep K Kalra, Ling Sun, Wenxiang Cong, and Ge Wang. 3-d convolutional encoder-decoder network for low-dose ct via transfer learning from a 2-d trained network. *IEEE transactions on medical imaging*, 37(6):1522–1534, 2018.

[109] Yu-Hsuan Shao, Kevin Tsai, Sinae Kim, Yu-Jen Wu, and Kitaw Demissie. Exposure to tomographic scans and cancer risks. *JNCI cancer spectrum*, 4(1):pkz072, 2020.

[110] Lin Shui, Haoyu Ren, Xi Yang, J. Li, Ziwei Chen, C. Yi, Hong Zhu, and Pixian Shui. The era of radiogenomics in precision medicine: An emerging approach to support diagnosis, treatment decisions, and prognostication in oncology. *Frontiers in Oncology*, 10, 2020.

[111] Emil Y Sidky and Xiaochuan Pan. Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. *Physics in Medicine & Biology*, 53(17):4777, 2008.

[112] Ethan A Smith, Jonathan R Dillman, Mitchell M Goodsitt, Emmanuel G Christodoulou, Nahid Keshavarzi, and Peter J Strouse. Model-based iterative reconstruction: effect on patient radiation dose and image quality in pediatric body ct. *Radiology*, 270(2):526–534, 2014.

[113] H Sujana, S Swarnamani, and S Suresh. Application of artificial neural networks for the classification of liver lesions by image texture parameters. *Ultrasound in medicine & biology*, 22(9):1177–1181, 1996.

[114] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[115] Sanna Suoranta, Kirsi Holli-Helenius, Päivi Koskenkorva, Eini Niskanen, Mervi Könönen, Marja Äikiä, Hannu Eskola, Reetta Kälviäinen, and Ritva Vanninen. 3d texture analysis reveals imperceptible mri textural alterations in the thalamus and putamen in progressive myoclonic epilepsy type 1, epm1. *PloS one*, 8(7):e69905, 2013.

[116] Richard N Sutton and Ernest Lenard Hall. Texture measures for automatic classification of pulmonary disease. *IEEE Transactions on Computers*, 100(7):667–676, 1972.

[117] Chao Tang, Jie Li, Linyuan Wang, Ziheng Li, Lingyun Jiang, Ailong Cai, Wenkun Zhang, Ningning Liang, Lei Li, and Bin Yan. Unpaired low-dose ct denoising network based on cycle-consistent generative adversarial network with prior image information. *Computational and mathematical methods in medicine*, 2019, 2019.

[118] Ryutaro Tanno, Daniel E Worrall, Aurobrata Ghosh, Enrico Kaden, Stamatios N Sotiropoulos, Antonio Criminisi, and Daniel C Alexander. Bayesian image quality transfer with cnns: exploring uncertainty in dmri super-resolution. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 611–619. Springer, 2017.

[119] Ryutaro Tanno, Daniel E Worrall, Enrico Kaden, Aurobrata Ghosh, Francesco Grussu, Alberto Bizzi, Stamatios N Sotiropoulos, Antonio Criminisi, and Daniel C Alexander. Uncertainty modelling in deep learning for safer neuroimage enhancement: Demonstration in diffusion mri. *NeuroImage*, 225:117366, 2021.

[120] National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409, 2011.

[121] Chunwei Tian, Lunke Fei, Wenxian Zheng, Yong Xu, Wangmeng Zuo, and Chia-Wen Lin. Deep learning on image denoising: An overview. *Neural Networks*, 2020.

[122] Sameer V Tipnis, Maria V Spampinato, John Hungerford, and Walter Huda. Thyroid doses and risks to adult patients undergoing neck ct examinations. *American Journal of Roentgenology*, 204(5):1064–1068, 2015.

[123] Florent Tixier, Catherine Cheze Le Rest, Mathieu Hatt, Nidal Albarghach, Olivier Pradier, Jean-Philippe Metges, Laurent Corcos, and Dimitris Visvikis. Intratumor heterogeneity characterized by textural features on baseline 18f-fdg pet images predicts response to concomitant radiochemotherapy in esophageal cancer. *Journal of Nuclear Medicine*, 52(3):369–378, 2011.

[124] Peter Toft. The radon transform. *Theory and Implementation (Ph. D. Dissertation)(Copenhagen: Technical University of Denmark)*, 1996.

[125] Alberto Traverso, Leonard Wee, Andre Dekker, and Robert Gillies. Repeatability and reproducibility of radiomic features: a systematic review. *International Journal of Radiation Oncology* Biology* Physics*, 102(4):1143–1158, 2018.

[126] JJM van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R.G.H Beets-Tan, J.C. Fillon-Robin, S. Pieper, and H.J.W.L Aerts. Computational radiomics system to decode the radiographic phenotype. cancer research. *Cancer Research*, 77(21):e104–e107, 2017.

[127] Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.

[128] J. Wang, H. Lu, T. Li, and Z. Liang. Sinogram noise reduction for low-dose ct by statistics-based nonlinear filters. *Medical Imaging*, 5747:2058–2066, Apr 2005.

[129] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[130] Stefanie Warnat-Herresthal, Hartmut Schultze, Krishnaprasad Lingadahalli Shastry, Sathyanarayanan Manamohan, Saikat Mukherjee, Vishesh Garg, Ravi Sarveswara, Kristian Händler, Peter Pickkers, N Ahmad Aziz, et al. Swarm learning for decentralized and confidential clinical machine learning. *Nature*, 594(7862):265–270, 2021.

[131] Arthur Robert Weeks, Lloyd J Sartor, and Harley R Myler. Histogram specification of 24-bit color images in the color difference (cy) color space. *Journal of electronic imaging*, 8(3):290–300, 1999.

[132] L. Wei, Y. Lin, and W. Hsu. Using a generative adversarial network for ct normalization and its impact on radiomic features. In *IEEE International Symposium on Biomedical Imaging*, Iowa City, USA, Apr 2020.

[133] Leihao Wei, Yannan Lin, and William Hsu. Using a generative adversarial network for ct normalization and its impact on radiomic features. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 844–848. IEEE, 2020.

[134] Martin J Willemink and Peter B Noël. The evolution of image reconstruction for ct—from filtered back projection to artificial intelligence. *European radiology*, 29(5):2185–2195, 2019.

[135] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum. Generative adversarial networks for noise reduction in low-dose ct. *IEEE Transactions on Medical Imaging*, 36(12):2536–2545, Dec 2017.

[136] Jelmer M Wolterink, Tim Leiner, Max A Viergever, and Ivana Išgum. Generative adversarial networks for noise reduction in low-dose ct. *IEEE transactions on medical imaging*, 36(12):2536–2545, 2017.

[137] Qiong Xu, Hengyong Yu, Xuanqin Mou, Lei Zhang, Jiang Hsieh, and Ge Wang. Low-dose x-ray ct reconstruction via dictionary learning. *IEEE transactions on medical imaging*, 31(9):1682–1697, 2012.

[138] Jinzhong Yang, Lifei Zhang, Xenia J Fave, David V Fried, Francesco C Stingo, Chaan S Ng, and Laurence E Court. Uncertainty analysis of quantitative imaging features extracted from contrast-enhanced ct in lung tumors. *Computerized Medical Imaging and Graphics*, 48:1–8, 2016.

[139] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang. 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, June 2018.

[140] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M.K. Kalra, Y. Zhang, L. Sun, and G. Wang. 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, Jun 2018.

[141] X. Yi and P. Babyn. Sharpness-aware low-dose ct denoising using conditional generative adversarial network. *Journal of Digit Imaging*, 31(31):655–669, Feb 2018.

[142] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552, 2019.

[143] C. You, G. Li, Y. Zhang, X. Zhang, H. Shan, M. Li, S. Ju, Z. Zhao, Z. Zhang, W. Cong, M. Vannier, P. Saha, E. Hoffman, and G. Wang. Ct super-resolution gan constrained by the identical, residual, and cycle learning ensemble (gan-circle). *IEEE Transactions on Medical Imaging*, 39(1):188–203, Jan 2020.

[144] Chenyu You, Guang Li, Yi Zhang, Xiaoliu Zhang, Hongming Shan, Mengzhou Li, Shenghong Ju, Zhen Zhao, Zhuiyang Zhang, Wenxiang Cong, et al. Ct super-resolution gan constrained by the identical, residual, and cycle learning ensemble (gan-circle). *IEEE Transactions on Medical Imaging*, 2019.

[145] S. Young, P. Lo, J. Hoffman, M. Wahi-Anwar, M. Brown, M. McNitt-Gray, and F. Noo. Th-ab-207a-05: A fully-automated pipeline for generating ct images across a range of doses and reconstruction methods. *Medical Physics*, 43(6), 6 2016.

[146] R. Zhang, P. Isola, A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[147] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[148] Binsheng Zhao, Yongqiang Tan, Wei-Yann Tsai, Jing Qi, Chuanmiao Xie, Lin Lu, and Lawrence H Schwartz. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Scientific reports*, 6(1):1–7, 2016.

[149] Can Zhao, Aaron Carass, Blake E Dewey, and Jerry L Prince. Self super-resolution for magnetic resonance images using deep networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 365–368. IEEE, 2018.

[150] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for neural networks for image processing. *arXiv preprint arXiv:1511.08861*, 2015.

[151] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[152] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 465–476. Curran Associates, Inc., 2017.