Dissertation

submitted to the
Combined Faculty of Mathematics, Engineering and Natural Sciences
of Heidelberg University, Germany
for the degree of
Doctor of Natural Sciences

Put forward by

Elias Eulig

born in: Hanover, Germany
Oral examination: 07.05.2025

Challenges and Opportunities of Deep Learning in X-Ray Imaging and Computed Tomography

Referees: Prof. Dr. Jürgen Hesser
Prof. Dr. Marc Kachelrieß

Abstract

Deep learning has revolutionized medical imaging by providing state-of-the-art solutions for a wide range of tasks. However, the use of deep learning in medical imaging also comes with its own set of challenges, three of which are addressed in the projects presented in this cumulative thesis.

The first project focuses on the fair evaluation of deep learning-based low-dose computed tomography (LDCT) image denoising algorithms by introducing a novel benchmark framework. The second project addresses the interpretability and robustness of deep learning algorithms for LDCT image denoising by investigating their invariances (i.e., which features in the images they learned to represent and which to ignore). The third project tackles the scarcity of data for deep learning-based digital subtraction angiography (DSA) by simulating paired training data.

The proposed methods are capable of overcoming the respective challenges associated with deep learning in medical imaging and through this could enable the development of better and safer algorithms for clinical practice.

◇ ◇ ◇

Zusammenfassung

Deep Learning revolutionierte die medizinische Bildgebung, da es modernste Lösungen für eine breite Palette von Aufgaben bietet. Der Einsatz von Deep Learning in der medizinischen Bildgebung bringt jedoch auch eine Reihe von Herausforderungen mit sich. Drei dieser Herausforderungen widmen sich die in dieser kumulativen Arbeit vorgestellten Projekte.

Das erste Projekt konzentriert sich auf die faire Evaluation von Deep-Learning-basierten Algorithmen zur Entrauschung von Niedrigdosis-CT-Bildern durch die Entwicklung eines neuen Benchmarks. Das zweite Projekt befasst sich mit der Interpretierbarkeit und Robustheit von Deep-Learning-Algorithmen für die Entrauschung von Niedrigdosis-CT-Bildern durch die Untersuchung ihrer Invarianzen (d.h. welche Merkmale in den Bildern sie ignorieren und welche nicht). Das dritte Projekt adressiert die Limitiertheit von Daten für die Deep Learning-basierte digitale Subtraktionsangiographie (DSA), indem gepaarte Daten für das Training von Deep-Learning-Algorithmen simuliert werden.

Die vorgeschlagenen Methoden sind in der Lage, die jeweiligen Herausforderungen zu überwinden, die mit Deep Learning in der medizinischen Bildgebung verbunden sind, und könnten dadurch die Entwicklung besserer und sichererer Algorithmen für die klinische Praxis ermöglichen.

Acknowledgements

I want to thank Prof. Dr. Jürgen Hesser for his role as first examiner and his support and guidance throughout my academic journey, spanning from my bachelor's thesis to the present work.

◇◇◇

I want to thank Prof. Dr. Marc Kachelrieß for letting me work in his department and for the scientific supervision of this thesis. Marc motivated me throughout the whole time, provided innumerable ideas and suggestions, and was always open for discussions. Without his support, this work would not have been possible. Thank you!

◇◇◇

I want to thank Prof. Dr. Björn Ommer and Prof. Dr. Joao Seco for inspiring discussions and great advice throughout this work and for being part of my thesis advisory committee.

◇◇◇

I want to thank Prof. Dr. Loredana Gastaldo and Prof. Dr. Tilman Plehn for agreeing to serve as examiners for the oral defense.

◇◇◇

I want to thank all supervisors, mentors, and managers throughout my academic journey who taught me how to conduct scientific research. These include (in chronological and alphabetical order) Prof. Dr. Moritz Helmstaedter, Dr. Alessandro Motta, Dr. Joscha Maier, Dr. Adam Wang, Dr. Volker Fischer, Dr. Patrick Blöbaum, and Dr. Dominik Janzing.

◇◇◇

I want to thank all current and former colleagues in the Department of X-Ray Imaging and Computed Tomography at the German Cancer Research Center for inspiring discussions, powerful bouldering sessions, and a wonderful atmosphere.

◇◇◇

I am deeply grateful to my friends and family for their encouragement, support, and love.

Contents

List of Abbreviations

ALARA As low as reasonably achievable

CAD Computer-aided diagnosis

CBCT Cone-beam computed tomography

cINN Conditional invertible neural network

CNR Contrast-to-noise ratio

CT Computed tomography

DML Deep metric learning

DNN Deep neural network

DSA Digital subtraction angiography

DSC Dice-Sørensen coefficient

FBP Filtered backprojection

IQA Image quality assessment

IR Iterative reconstruction

L-system Lindenmayer system

LDCT Low-dose computed tomography

LLM Large language model

MAR Metal artifact reduction

MC Monte Carlo

ML Machine learning

MRI Magnetic resonance imaging

OOD Out-of-distribution

PET Positron emission tomography

PSNR Peak signal-to-noise ratio

RMSE Root mean squared error

SSIM Structural similarity index measure

VAE Variational autoencoder

VIF Visual information fidelity

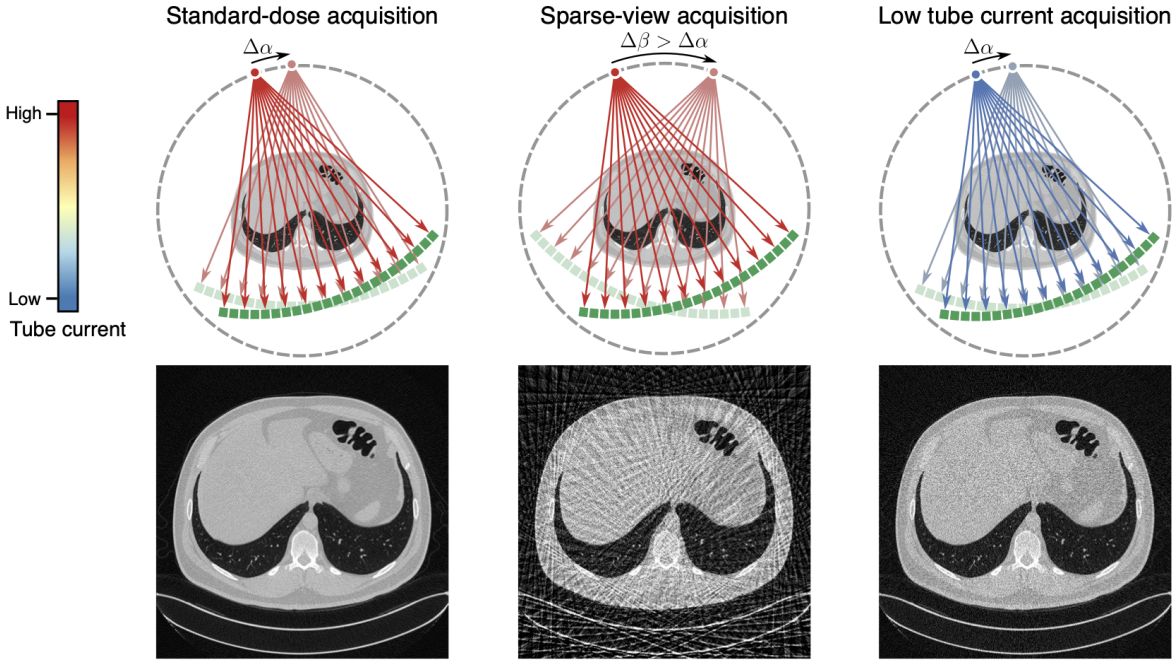
Chapter 1

Introduction

Computed tomography (CT) is widely regarded as the workhorse of modern radiology [1] with indispensable applications in a wide range of clinical scenarios. Its origins trace back to the discovery of X-rays by Wilhelm Conrad Röntgen in 1895, who was also the first one to realize their potential for medical imaging [2]. Only months after Röntgen acquired the famous first medical X-ray image of his wife's hand, X-rays were used for diagnostic and interventional purposes in clinical settings. Six years later, in 1901, Röntgen was awarded the first Nobel Prize in Physics for his discovery. Today, X-ray images –and continuous acquisitions thereof, referred to as X-ray fluoroscopy– remain one of the most frequently used medical imaging modalities [3, 4].

However, in many applications, the diagnostic value of X-ray images is limited by the fact that they only provide two-dimensional projections of the three-dimensional anatomy. This limitation was overcome in 1972 with the development of the first CT scanner by Sir Godfrey Hounsfield which enabled the acquisition of cross-sectional images of the human body [5, 6]. Since then, numerous technological advancements have been made to improve the speed, resolution, and image quality of CT scanners. Today, CT is used for diagnostic imaging of innumerable diseases covering all parts of the body, including diagnosis of cancer [7, 8], cardiovascular diseases [9], and trauma [10]. Beyond diagnostics, flat-panel cone-beam computed tomography (CBCT) is also used for planning and guiding surgical and minimally-invasive procedures such as vascular stenting, chemo- and radioembolization [11, 12, 13, 14, 15], as well as for image guidance in radiation therapy [16].

Despite their widespread use, X-ray imaging and CT have a major drawback, the use of ionizing radiation which can have deleterious effects in humans, including the induction of cancer. Although, risks associated with very low doses (i.e., $< 10\,\mathrm{mSv}$) of radiation remain a topic of debate [17, 18, 19, 20, 21], it is generally accepted that the risk of radiation-induced cancer increases with the dose of radiation, and thus clinical CT scans must follow the ALARA (as low as reasonably achievable) principle [22, 23], making dose reduction one of the primary research areas in the field. Two of the (conceptually) most straightforward approaches to low-dose computed tomography (LDCT) are by means of lowering the tube current or by reducing



▲ **Figure 1.1.** Illustration of two conceptually straightforward approaches to low-dose CT. In sparse-view CT (center) the number of angular positions from which X-ray projections are acquired is reduced (hence increasing the angular increment between two neighboring projections). Alternatively, one may decrease the tube current while still acquiring the full set of projections (right). Compared to a standard-dose acquisition (left), both approaches lead to artifacts in the reconstructed images (bottom row) which can impair their diagnostic value. $C = -300\,\text{HU}$, $W = 1500\,\text{HU}$.

the number of angular positions from which X-ray projections are acquired, referred to as sparse-view CT[1] (Fig. 1.1). However, both approaches lead to artifacts in the reconstructed images, which can impair their diagnostic value in many clinical settings [24, 25].

To mitigate these and other artifacts (e.g., caused by metal or scatter) in CT a variety of classical algorithms have been developed. Most notable among these are iterative reconstruction (IR) algorithms, which outperform analytical filtered backprojection (FBP) by incorporating prior knowledge into an iterative reconstruction loop. This prior knowledge can be about e.g., the acquisition process [26, 27, 28, 29, 30] or the object that is being imaged [31, 32]. However, these algorithms are often computationally expensive, and require careful hyperparameter selection, limiting their use in clinical practice.

With the advent of deep learning, a new class of algorithms for CT image reconstruction has emerged that revolutionized the field. These algorithms can broadly be categorized into three groups: The first two groups consist of algorithms that perform some pre- or post-processing using deep neural networks (DNNs) to reduce artifacts in raw-data (sinogram) domain or image domain, while still relying on classical algorithms for the main reconstruction task [33, 34, 35, 36, 37, 38, 39, 40]. The third group consists of algorithms that perform the main reconstruction task

[1] Note that sparse-view acquisitions are not straightforward to implement in practice and are not yet available in commercial CT scanners.

entirely using some DNN [41, 42, 43]. Besides these, DNNs have been used to optimize multiple other aspects of the imaging pipeline such as image registration [44], dose estimation [45, 46, 47], or tube current modulation [48, 49]. However, the use of DNNs in medical imaging also comes with its own set of challenges, three of which will be discussed in the following.

Scarcity of data: Due to their large number of parameters, modern DNNs require large amounts of (usually paired) data to be trained effectively. However, in the medical domain, such data are often scarce due to privacy concerns, the rarity of certain conditions, or the high cost of annotations [50]. In medical imaging, this problem is further exacerbated, as paired data, with and without artifacts, are rarely available at all. E.g., in LDCT, patients are typically not allowed to undergo multiple scans at different dose levels, and thus for the training of LDCT denoising networks on paired data, low-dose acquisitions must be simulated from standard-dose scans instead [51, 52] (as done for the LDCT reconstructions shown in Fig. 1.1). Similarly, for the task of metal artifact reduction (MAR), metal and corresponding artifacts must be added to scans without metal implants via simulation to create paired datasets for training [53, 54].

Interpretability and robustness: Predictions made by DNNs are notoriously difficult to interpret due to their complexity and thus lack of decomposability into intuitively understandable components [55]. This can clearly be problematic when employed in clinical settings but also hinder research, where identification of potential failure modes is crucial for the development of better algorithms [56]. Interpretability can also be seen as a prerequisite for trust in an algorithm's ability to generalize to real-world data, a property particularly relevant for medical imaging, where the use of simulated data is common due to the scarcity of (paired) clinical data (see previous paragraph).

Fair evaluation of models: Another challenge lies in the fair and meaningful evaluation of DNNs for medical imaging, which is a threefold problem. Firstly, the lack of ground truth for many clinical tasks makes quantitative evaluation difficult or impossible, forcing researchers to rely on qualitative assessment instead. Secondly, many existing quantitative measures for image quality assessment (IQA) are not in agreement with human readers, which are considered the gold standard for IQA of medical images [57, 58, 59, 60]. However, conducting such reader studies is expensive and time-consuming, and thus often not feasible [61]. Even if feasible, it remains subjective and prevents the comparison of algorithms across different studies [62]. Thirdly, the lack of standardized openly available data for many tasks due to privacy concerns or proprietary interests makes it often difficult to reproduce results and compare algorithms across different studies even if open-source code is available.

The three projects presented in this thesis address each of the previously described challenges. The first project (Chapter 2) focuses on the fair evaluation of DNN-based LDCT image denoising algorithms by introducing a novel benchmark framework. The second project (Chapter 3) addresses the interpretability and robustness of DNNs for LDCT image denoising by investigating their invariances (i.e., what they learned to represent and what to ignore). The third project (Chapter 4) tackles the scarcity of data for deep learning-based digital subtraction angiography (DSA) by simulating paired training data.

The structure of this cumulative thesis is as follows: In the following chapters, the three projects are presented (Chapters 2 to 4). For each project we first provide a brief summary of the work, followed by the respective paper for the interested reader. We then discuss our projects and how they addressed the aforementioned challenges with DNNs for medical imaging in Chapter 5 and conclude with a summary in Chapter 6. In Appendix A to C, we also provide the supplementary materials of each of the papers.

Chapter 2

Benchmarking Low-Dose CT Denoising Networks

E. Eulig, B. Ommer, and M. Kachelrieß. "Benchmarking Deep Learning-Based Low-Dose CT Image Denoising Algorithms". In: *Medical Physics* 51.12 (2024), pp. 8776–8788

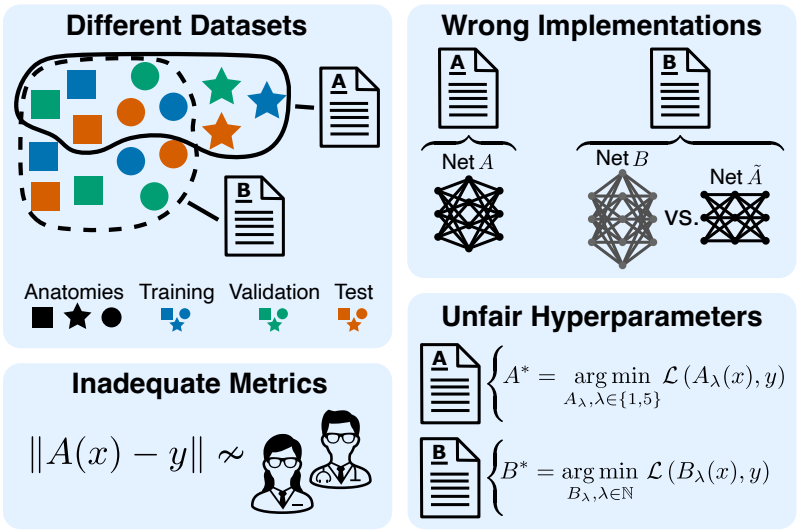
In this chapter our work on benchmarking deep learning-based LDCT image denoising algorithms is presented. We begin with a brief summary of our work in Sec. 2.1. The full paper [63] is then presented in Sec. 2.2.

2.1 Summary

2.1.1 Introduction

Reducing the dose and thus potential radiation risk associated with a CT exam is one of the primary research areas in the field (c.f. Chapter 1). While there exist many different hardware- and software-based approaches to this (e.g., automatic exposure control, tube current modulation, bowtie filters [64]), LDCT by means of reducing the tube current is probably the most straightforward approach. However, the variance of the noise in the reconstructed image is inversely proportional to the tube current (which in turn is proportional to the number of X-ray quanta measured at the detector) [65], thus potentially impairing the diagnostic value of the image. In recent years, deep learning-based algorithms to reduce the noise in such LDCT images showed great potential over classical algorithms due to superior image quality [24, 66] and possibly reduced computational burden. This led to a plethora of new algorithms being proposed, many of which claim to outperform the previous state-of-the-art.

In this work, we identified four main flaws in the experimental setup of many of these studies that limit the comparability and reproducibility of the results and their claimed improvements (Fig. 2.1):



► **Figure 2.1.** Overview of flaws that we identified in the experimental setup of many deep learning-based LDCT denoising methods. These flaws limit the comparability and reproducibility of the results and their claimed improvements. Figure taken from [63].

1. *Different datasets:* The lack of unified benchmark datasets (including unified training, validation, and test sets) makes it difficult to compare the performance of different algorithms. This also means that results presented in papers are often not reproducible as the training data (splits) remain unknown, even if the underlying dataset is publicly available.

2. *Unfair hyperparameters:* Most studies do not perform a rigorous hyperparameter[1] optimization, leading to possibly suboptimal performance of the comparison methods as well as of the proposed method. This is particularly problematic given the use of different datasets (see previous point), as the optimal hyperparameters of a method may vary between datasets.

3. *Missing open-source implementations:* The absence of open-source implementations for many algorithms often forces researchers to implement and validate the comparison methods themselves. This can lead to errors and biases, compromising the validity of experimental results.

4. *Inadequate evaluation metrics:* The performance of DNNs for LDCT image denoising is commonly evaluated using full-reference (objective) IQA metrics such as root mean squared error (RMSE), peak signal-to-noise ratio (PSNR), or structural similarity index measure (SSIM). However, these metrics typically do not correlate well with human readers, thus limiting their clinical relevance [58, 60, 59]. Very few studies include metrics that have been shown to better correlate with human readers, such as visual information fidelity (VIF) [67, 60, 59], or conduct reader studies to verify the results.

2.1.2 Methods

To overcome the aforementioned flaws, we proposed a benchmark framework for assessing deep learning-based LDCT denoising algorithms in a fair and reproducible manner. Our framework

[1] Hyperparameters typically refer to parameters of the training process (e.g., learning rate or mini-batch size) that are not learned during the training phase of the algorithm. Note that architectural choices (e.g., number of layers of a network) are also sometimes considered hyperparameters.

consists of the following components:

1. *Unified dataset:* For evaluation, we use the *Low Dose CT and Projection Dataset* [68], which includes 150 routine-dose scans of the head, chest, and abdomen. For each scan, the authors provide simulated low-dose scans by means of noise insertion in the projection domain. We split the dataset into training, validation, and test sets, and released code to download, preprocess, and split the data as part of our benchmark suite. This allows for a fair comparison of different algorithms on the same data.

2. *LDCT denoising algorithms:* In our benchmark we consider eight popular deep learning-based LDCT denoising algorithms that have been previously proposed. When open-source implementations were not available, we implemented and validated the algorithms ourselves and made the code publicly available.

3. *Hyperparameter optimization:* To ensure a fair comparison between the different algorithms we performed a rigorous hyperparameter optimization using sequential model-based optimization. In particular, for each algorithm, we first identified hyperparameters and suitable ranges and then performed a Bayesian optimization to maximize the SSIM on the validation set.

4. *Evaluation metrics:* We evaluated all algorithms using PSNR, SSIM, and VIF. Additionally, we introduced a novel metric to evaluate the preservation of radiomic features in the denoised images by measuring the similarity of radiomic features extracted from the denoised images to those extracted from the routine-dose scans. We also computed physical metrics such as contrast-to-noise ratio (CNR) and CT number accuracy to quantify the technical performance of the algorithms and evaluated the algorithms' ability to reconstruct lesions.

2.1.3 Results

Using our benchmark suite we evaluated the performance of the eight deep learning-based LDCT denoising algorithms. We found that most algorithms showed only marginal improvements over the past six years, with many algorithms performing similarly to RED-CNN [37], one of the earliest approaches. GAN-based models showed some superiority in preserving radiomic features, particularly in high-noise settings. However, the newer methods did not consistently outperform older techniques. The findings suggest a need for more rigorous validation and evaluation in LDCT denoising research, with the proposed benchmark providing a crucial foundation for future developments.

2.2 Paper: Benchmarking Deep Learning-Based Low-Dose CT Image Denoising Algorithms

The following pages contain the full paper [63]. To improve readability, the supplementary material is provided in Appendix A.

DOI: 10.1002/mp.17379

RESEARCH ARTICLE

MEDICAL PHYSICS

Benchmarking deep learning-based low-dose CT image denoising algorithms

Elias Eulig[1,2] | **Björn Ommer**[3] | **Marc Kachelrieß**[1,4]

[1]Division of X-Ray Imaging and Computed Tomography, German Cancer Research Center (DKFZ), Heidelberg, Germany

[2]Faculty of Physics and Astronomy, Heidelberg University, Heidelberg, Germany

[3]CompVis @ LMU Munich, MCML, Munich, Germany

[4]Medical Faculty Heidelberg, Heidelberg University, Heidelberg, Germany

Correspondence
Elias Eulig, Division of X-Ray Imaging and Computed Tomography, German Cancer Research Center (DKFZ), Heidelberg, Germany.
Email: elias.eulig@dkfz.de

Funding information
Helmholtz International Graduate School for Cancer Research

Abstract

Background: Long-lasting efforts have been made to reduce radiation dose and thus the potential radiation risk to the patient for computed tomography (CT) acquisitions without severe deterioration of image quality. To this end, various techniques have been employed over the years including iterative reconstruction methods and noise reduction algorithms.

Purpose: Recently, deep learning-based methods for noise reduction became increasingly popular and a multitude of papers claim ever improving performance both quantitatively and qualitatively. However, the lack of a standardized benchmark setup and inconsistencies in experimental design across studies hinder the verifiability and reproducibility of reported results.

Methods: In this study, we propose a benchmark setup to overcome those flaws and improve reproducibility and verifiability of experimental results in the field. We perform a comprehensive and fair evaluation of several state-of-the-art methods using this standardized setup.

Results: Our evaluation reveals that most deep learning-based methods show statistically similar performance, and improvements over the past years have been marginal at best.

Conclusions: This study highlights the need for a more rigorous and fair evaluation of novel deep learning-based methods for low-dose CT image denoising. Our benchmark setup is a first and important step towards this direction and can be used by future researchers to evaluate their algorithms.

KEYWORDS
benchmarking, computed tomography, deep learning, denoising, low-dose

1 | INTRODUCTION

Computed tomography (CT) is an important imaging modality, with numerous applications including biology, medicine, and nondestructive testing. However, the use of ionizing radiation remains a key concern and thus clinical CT scans must follow the ALARA (as low as reasonably achievable) principle.[1,2] Therefore, reducing the dose and thus radiation risk is of utmost importance and one of the primary research areas in the field.

A straightforward approach to reduce dose is by lowering the tube current (i.e., reducing the x-ray intensity). However, this comes at the cost of deteriorated image quality due to increased image noise and thus potentially reduced diagnostic value. To alleviate this drawback, numerous algorithms have been proposed to solve the task of low-dose CT (LDCT) denoising, that is, reducing image noise in the reconstructed image (or volume).

Iterative reconstruction (IR) techniques incorporate prior knowledge in the reconstruction process and

then update the reconstructed image iteratively. The prior knowledge may model statistical properties of the noise,[3] properties of the object to be reconstructed,[4] or parameters of the CT system. While IR techniques can be used to reduce numerous other artifacts compared to conventional filtered back projection (FBP), they are computationally expensive, which limits their clinical applicability. On the other hand, filtering techniques to reduce noise are fast and easy to implement into various reconstruction frameworks. The filtering may either be performed in projection domain, image domain, or both, and using a wide range of algorithms.[5–9] Recently deep learning-based filtering, particularly in the image domain, became increasingly popular.[10–22] The majority of the proposed methods learn a mapping from low-dose images to high-dose images in a supervised fashion using a deep neural network (DNN). Of the numerous proposed methods, most suggestions for improvement alter the network structure, loss function, or training strategy. Publications often claim ever improving performance which is commonly demonstrated by improved image quality metrics (e.g., peak signal-to-noise ratio, structural similarity) in experiments on simulated or clinical data.

In this work, we identify several flaws in the experimental setup of such methods which limit the verifiability of the claimed improvements. These include the lack of a common benchmark dataset, the use of inadequate metrics with little relation to diagnostic value, and unfair choice of hyperparameters for reference methods. Reproducibility and verifiability of scientific results, however, are paramount to scientific advancements of a field, and thus efforts towards fair benchmarking of existing and future algorithms are of utmost importance. To this end, we make the following contributions:

1. We identify multiple flaws in the experimental setup of previously proposed methods which hinder the verifiability of their claimed improvements.
2. We propose a benchmark setup[1] for deep learning-based LDCT denoising methods, which aims to overcome those flaws and allows for a fair evaluation of existing algorithms and those yet to come.
3. In a comprehensive and fair evaluation of several existing algorithms we find that there has been little progress over the past six years and many of the newer methods perform statistically similar or worse compared to older ones.

2 | RELATED WORK

In this section, we review existing works on deep learning-based LDCT denoising and image quality assessment (IQA) of medical images.

[1] https://github.com/eeulig/ldct-benchmark

2.1 | Deep learning-based LDCT denoising

CT image reconstruction aims at solving the linear system $\boldsymbol{R}x = p$, with $p \in \mathbb{R}^M$ denoting the measurements in projection domain, $x \in \mathbb{R}^N$ being the volume to be reconstructed, and $\boldsymbol{R} \in \mathbb{R}^{M \times N}$ the Radon transform. LDCT generally aims at reconstructing x using less dose, which can be for example, accomplished by lowering the tube current, thus increasing the noise in p and x, or by lowering the number of measurements M, leading to sparse-view artifacts in x. Since previous studies indicate that DNN-based correction of the former can be superior, we here consider the task of LDCT denoising.[23] Based on the domain (p, x, or both) in which they operate, deep learning-based methods for LDCT image denoising can be divided into three categories: projection-domain, image-domain, and dual-domain.

Projection-domain methods aim to learn a mapping $f_\theta : p' \rightarrow p$ from low-dose projections p' to high-dose projections p, where f_{θ^*} is realized by a DNN, parameterized by weights θ. These weights are either optimized in a supervised setting via

$$\theta^* = \arg\min_\theta \mathbb{E}_{p', p \sim \mathcal{D}^{\text{train}}} \| f_\theta(p') - p \| \,, \tag{1}$$

with $\| \cdot \|$ being some norm,[24,25] or unsupervised, exploiting structural similarities between adjacent projections.[26,27] The denoised projections can then be reconstructed using either of the standard reconstruction techniques.[28–30]

Image-domain methods aim to directly learn a mapping $g_\phi : x' \rightarrow x$ from low-dose images x' (i.e., images reconstructed from low-dose projections p' using FBP) to high-dose images x. Similar to Equation (1), weights are typically optimized in a supervised setting, where the mean-squared error (MSE), or some other pixel- or feature-based loss between prediction and high-dose image x is minimized,[10–14,17,19,21,22,31] or g is trained together with a discriminator as a generative adversarial network (GAN).[15,18,20] Notable other works investigate unsupervised- or self-supervised training strategies, or leverage the intrinsic image prior of DNNs.[32]

Lastly, dual-domain methods operate in both domains x and p simultaneously, by employing two separate networks f and g, respectively. Networks are trained either separately using aforementioned loss functions[33,34] or in an end-to-end fashion using a differentiable analytical reconstruction layer.[35–37]

In this work we focus on image-domain methods which dominate the research field. This is mainly due to the abundance of open source datasets, where paired high- and low-dose images are readily available.[38,39] In contrast, projection data are generally proprietary and thus difficult to access.[40] The few datasets that provide

them usually do so only for a (vendor-specific) subset of the data and handling of them can be cumbersome due to (hidden) preprocessing steps in the reconstruction pipeline of the vendor.[39,41] Many of the principles in the design of our benchmark setup, however, can be applied to the evaluation of projection-domain or dual-domain methods as well.

2.2 | Medical image quality assessment

Common full-reference quantitative measures for natural image quality assessment include the structural similarity index measure (SSIM)[42] and peak signal-to-noise ratio (PSNR). However, these metrics are usually not in agreement with human readers, which are considered the gold standard for image quality assessment of medical images.[43–45] These are conducted by measuring the accuracy of multiple radiologists when performing some task (e.g., lesion detection or segmentation) using certain images. However, this metric relies, and is dependent on the definition of a suitable task. Therefore, the subjective assessment of overall diagnostic quality by radiologists is a common alternative measure.[46] Nonetheless, since conducting multiple-reader studies is time-consuming and expensive, most algorithms for enhancement of medical images are still evaluated using quantitative metrics such as SSIM or PSNR.

In refs. [45, 46], the authors find that multiple other metrics, including the visual information fidelity (VIF),[47] have higher correlation with human reader ratings compared to SSIM and PSNR for both CT and magnetic resonance (MR) images. Furthermore, notable recent works investigate the use of radiomic features to provide a clinically meaningful measure for the quality of medical images without the drawbacks of human reader studies.[48–50]

Moreover, many physical image quality metrics for evaluating different aspects of the technical performance of CT equipment exist such as the modulation transfer function (MTF), contrast-to-noise ratio (CNR), noise power spectrum (NPS), and CT number accuracy. However, these quantities often rest on strong assumptions about the imaging system and reconstruction algorithm such as linearity, shift-invariance, or stationarity of the noise, many of which are violated for IR or deep learning-based reconstruction methods.[51–54] Another drawback is that these metrics are commonly evaluated using phantom measurements, thus posing an out-of-distribution problem for deep learning-based methods which are trained exclusively on clinical data. Even for methods that are trained on a mixture of phantom and patient data (e.g., GE Healthcare's TrueFidelity™[55]), results from phantom measurements may not be representative of the performance on clinical data.

3 | FLAWS OF CURRENT EVALUATION PROTOCOLS

In this section we will outline the main problems with current evaluation protocols for deep learning-based image-domain LDCT denoising (see Figure 1 for an overview).

3.1 | Different datasets

Unlike in many other disciplines of computer vision, particularly image denoising of natural images,[56–60] there exist no consensus regarding benchmark datasets for LDCT denoising. While most methods are trained and evaluated on the dataset provided as part of the *2016 NIHAAPM-Mayo Clinic LDCT Grand Challenge*[38] or the subsequently released (significantly larger both in number of images and anatomical sites) *LDCT and Projection data*,[39] authors of each method employ their own training, validation, and test split. Therefore, reported metrics across publications are not comparable. This is further exacerbated by the fact that performance of individual methods differs significantly between different anatomical sites and images (i.e., axial slices), as shown by our experiments.

3.2 | Unfair choice of hyperparameters

Very few publications on LDCT denoising methods report the application of hyperparameter optimization[61–63] for their own or the considered comparison methods. In none of the respective publications of the algorithms considered in this study, exhaustive hyperparameter optimization is performed. The 3/8 algorithms that report some form of hyperparameter optimization limit it to a grid search with few points over

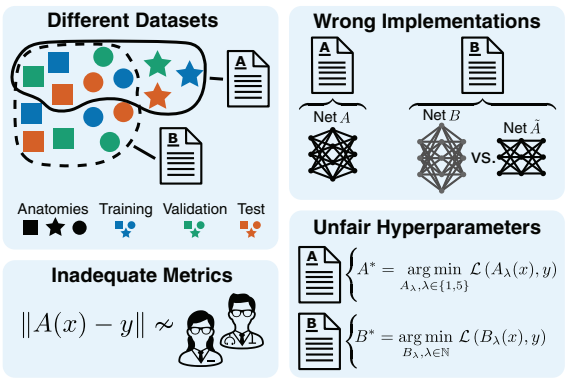


FIGURE 1 Overview of flaws in the experimental setup of many deep learning-based LDCT denoising methods, that limit their verifiability.

a single parameter (learning rate),[13,19] a subset of the comparison methods,[13] or their own method.[15] Often, authors simply use the hyperparameters reported in the reference publications.[12,13,15] This is particularly problematic given the choice of different datasets (cf. Section 3.1), where hyperparameters optimized by authors of method A on dataset $\mathcal{D}_A$ may not be optimal for the dataset $\mathcal{D}_B$ employed by authors B in their experiments.

3.3 | Missing open source implementations

With many authors not providing open source implementations of their algorithms, researchers are often left to implement comparison methods themselves. This increases the chances of errors.[64] Additionally, changing other aspects (such as the architecture of comparison methods[13]) can further bias experimental results.

3.4 | Inadequate metrics

Most LDCT denoising methods are evaluated using SSIM,[42] PSNR, or root-mean-square error (RMSE). While these are common metrics to quantify performance for natural image denoising, they are usually not in agreement with human readers for medical images (cf. Section 2.2), making it difficult to assess the extent to which the reported improvements actually translate into clinical benefits. This could be improved by the use of quantitative measures that are more suited for medical images (e.g., VIF), or experiments using human reader studies. In the respective publications of the eight algorithms considered in this study, however, most are evaluated using SSIM, RMSE, and PSNR only. Better metrics such as VIF or reader studies are employed in three publications only.

4 | BENCHMARK SETUP

In the following we present a benchmark setup to overcome the flaws of current evaluation protocols, as outlined in Section 3 that allows for a fair and clinically meaningful evaluation of DNNs for LDCT denoising.

4.1 | Dataset

For our benchmark setup we utilize the publicly available *Low dose CT and Projection Dataset*,[39] comprising a total of 150 scans of abdomen, head, and chest, (50 scans for each exam type) at routine dose levels. For each scan, simulated low-dose reconstructions (by means of noise insertion in the projection domain) at 25% dose for abdomen/head and 10% dose for chest, are available. For each exam type separately, data are split in 70%/20%/10% training/validation/test set and then linearly normalized to have zero mean, unit variance. Future studies might consider treating data normalization as an additional hyperparameter.[65] During training, we employ a weighted sampling scheme such that slices from each exam type and patient are sampled with equal probability. During testing, we reduce each scan to axial regions where the brain is present (for head scans), the lung is present (for chest scans), or the lung is not present (abdomen). We did not apply any data augmentation to the training data as we did not observe overfitting in our experiments for any of the methods. The code to reproduce exact dataset splits and all preprocessing is included in our benchmark suite.

4.2 | LDCT denoising algorithms

We consider eight DNN-based LDCT denoising algorithms proposed in the literature over the past six years. In the following we briefly describe each of the methods and refer the reader to the respective publications for more details. **CNN-10**[10] is a simple three layer CNN, trained to minimize the MSE between network output and high-dose targets. **RED-CNN**[12] and **ResNet**[31] are trained in the same fashion but employ deeper network architectures with residual connections compared to CNN-10. **WGAN-VGG**[20] and **DU-GAN**[15] are trained in an adversarial fashion,[66,67] where DU-GAN utilizes a U-net based discriminator.[68] **QAE**[13] is based on RED-CNN in both network architecture and training scheme, but employs quadratic convolutions. **TransCT**[22] is based on transformer blocks and also trained with an MSE loss. **Bilateral**[19] uses a trainable bilateral filter instead of a DNN, and thus substantially reduces the amount of free model parameters.

In Appendix A, we provide details on our implementation and verification of each of the algorithms.

4.3 | Hyperparameter optimization

As discussed in Section 3.2, for none of the methods a rigorous hyperparameter optimization was employed in the original publications. To ensure a fair comparison between different algorithms we optimize hyperparameters as follows. For each method we first identify hyperparameters and their suitable ranges. This includes general parameters such as learning rate, mini-batch size, patchsize and number of iterations, but also weighting factors in the loss functions (e.g., to balance adversarial and pixelwise loss in a GAN setting). Suitable ranges were determined from the respective papers (with sufficient margin) and whenever two methods had the

TABLE 1 Hyperparameters for all deep-learning based LDCT denoising methods considered in this study.

	Parameter	Prior
All algorithms	Learning rate	$\log \mathcal{U}(1\times10^{-5}, 0.01)$
	Maximum iterations	$\mathcal{U}(1\times10^{3}, 1\times10^{5})$
	Mini-batch size	$\mathcal{U}(2, 128)$
CNN-10[10]	Patchsize	$\mathcal{U}(32, 128)$
RED-CNN[12]	Patchsize	$\mathcal{U}(32, 128)$
WGAN-VGG[20]	β_1 of Adam	$\mathcal{U}(0.3, 0.9)$
	Loss weight: $\lambda_{\text{perceptual}}$	$\mathcal{U}(0, 1)$
	Critic updates	$\mathcal{U}(1, 5)$
	Patchsize	$\mathcal{U}(32, 128)$
ResNet[31]	Patchsize	$\mathcal{U}(32, 128)$
QAE[13]	Patchsize	$\mathcal{U}(32, 128)$
DU-GAN[15]	β_1 of Adam	$\mathcal{U}(0.3, 0.9)$
	Cutmix warmup	$\mathcal{U}(0, 1\times10^{4})$
	Loss weight: λ_{adv}	$\mathcal{U}(0, 1)$
	Loss weight: λ_{CM}	$\mathcal{U}(0, 10)$
	Loss weight: $\lambda_{\text{px,grad}}$	$\mathcal{U}(0, 40)$
	Critic updates	$\mathcal{U}(1, 5)$
	Patchsize	$\mathcal{U}(32, 128)$
TransCT[22]	—	—
Bilateral[19]	Learning rate for σ_r	$\log \mathcal{U}(1\times10^{-5}, 0.01)$
	Patchsize	$\mathcal{U}(32, 128)$
	Initalization for σ_r	$\mathcal{U}(0, 1)$
	Initalization for $\sigma_{x,y}$	$\mathcal{U}(0, 1)$

Note: The first three parameters are optimized for all algorithms (separately). Abbreviations: $\mathcal{U}$: uniform distribution; $\log \mathcal{U}$: log-uniform distribution.

same hyperparameter (e.g., learning rate or patchsize), we kept the prior distribution over the search space the same. All hyperparameters and their respective prior distributions are reported in Table 1. For each method, we then performed a black box hyperparameter tuning using sequential-model based optimization (SMBO). Such an automatic approach is preferred over manual (human) optimization as it avoids any potential bias by the practitioner, thus ensuring fair comparison of different models. Furthermore, SMBO has been shown to outperform both human optimization and non-sequential optimization schemes like grid search or random search on a variety of DNN and dataset combinations.[61,63]

Let $t_\lambda : \{f_\theta, \mathcal{D}^{\text{train}}, \lambda\} \rightarrow \theta^*$ denote the outcome of some training run of network f on training data $\mathcal{D}^{\text{train}}$ using hyperparameters λ. The aim of hyperparameter optimization is to find an optimial set of hyperparameters λ^*, that is,

$$\begin{aligned}\lambda^* &= \arg\max_\lambda \mathbb{E}_{x,y\sim\mathcal{D}^{\text{val}}}\left[M\left(y, f_{t_\lambda}(x)\right)\right] \\ &= \arg\max_\lambda \Psi(\lambda)\,, \end{aligned} \tag{2}$$

where M is some metric and $\mathcal{D}^{\text{val}}$ the validation dataset (not used during t_λ). Since evaluating $\Psi(\lambda)$ is expensive, requiring a full training run t_λ, one uses a probabilistic model p_Ψ, here constructed via Gaussian processes, as a surrogate for Ψ. For each iteration in the optimization process, we then find the most promising next point λ, to run the costly evaluation $\Psi(\lambda)$ for, by maximizing some acquisition function. In our experiments we used the expected improvement (EI)[61] as acquisition function:

$$\text{EI}(\lambda, \Psi^*) = \int \max(z - \Psi^*, 0)\, p_\Psi(z|\lambda) dz\,, \tag{3}$$

where Ψ^* refers to the expectation of M on the validation data for the best set of hyperparameters found so far (i.e., the one that maximizes the r.h.s. of Equation 2 up to now). As metric M that is optimized by the hyperparameter optimization, we used the SSIM for all networks. Optimizing the SSIM is favorable over other measures, since it is fast to compute, unlike for example, VIF, and not directly involved in the training process t_λ of any of the methods considered in this study (unlike e.g., RMSE). Further, note that for methods using a vanilla GAN loss, for example, ref. [15], simply minimizing the validation loss would not be suitable as it is not directly related to training progress. For each method, we perform 50 iterations of SMBO, sufficient to ensure convergence for all algorithms, as verified by our experiments.

After an optimal set of hyperparameters λ^* was found, we retrained a method using λ^* 10 times with different random seeds. If not stated otherwise, all reported standard deviations and significance tests (to compare two methods) are computed over those 10 training runs.

4.4 | Metrics

We evaluate all methods on the same test set comprising a total of 15 scans (5 head/chest/abdomen) using three common full-reference measures of image quality: SSIM, PSNR[2], and VIF. As described in Section 3.4, both SSIM and PSNR are common metrics to evaluate DNNs for LDCT denoising. We include VIF, since it has been shown to have higher correlation with human readers for medical images.[45,46]

Conducting human reader studies is time-consuming and expensive and would render the application of the proposed benchmark setup to future algorithms impossible. To nevertheless evaluate the algorithms in terms of clinically relevant image properties, we include an analysis of radiomic features. To this end, we compare the similarity of radiomic features extracted on the denoised images to those extracted on the high-dose image.

[2] We here omit evaluation of RMSE since it is related to the PSNR via PSNR = $20\log_{10}(I_{\text{max}}/\text{RMSE})$, with I_{max} being the maximum pixel value.

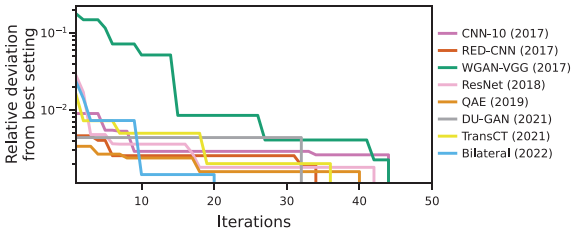


FIGURE 2 Evolution of relative deviation from best setting over the 50 iterations of Bayesian hyperparameter optimization. For each iteration i we show the relative deviation of the best network up to i from the final best configuration of hyperparameters (over all 50 iterations).

Definition 1 (Radiomic feature similarity). Let $\cos(x, y)$ be the cosine similarity between two vectors x and y:

$$\cos(x, y) = \frac{x \cdot y}{\|x\| \|y\|} . \tag{4}$$

Further, let $A = \{0, 1, 2, \dots, n\}$, with n being the number of algorithms considered, and index 0 being associated with the high-dose image. We denote with $R_{i,j}^{(s)}$ the radiomic feature $j \in \{1, 2, \dots, J\}$ extracted on scan s associated with algorithm $i \in A$. In order to get a task-agnostic metric, we assign an equal a-priori importance to each feature by normalizing

$$\tilde{R}_{i,j}^{(s)} = \frac{R_{i,j}^{(s)} - \max_{k \in A} R_{k,j}^{(s)}}{\max_{k \in A} R_{k,j}^{(s)} - \min_{k \in A} R_{k,j}^{(s)}} . \tag{5}$$

The radiomic feature similarity $\mathrm{RFS}_i^{(s)}$ of algorithm $i = 1, \dots, n$ on some scan s is then given as

$$\mathrm{RFS}_i^{(s)} = \cos\left(r_i^{(s)}, r_0^{(s)}\right), \quad r_i^{(s)} = \left(\tilde{R}_{i,1}^{(s)}, \dots \tilde{R}_{i,J}^{(s)}\right) . \tag{6}$$

Radiomic features are commonly extracted on segmentations of tumors or entire organs. On the high-dose scans of the test data, we therefore segment the following organs using the TotalSegmentator[69]: lung on chest scans, liver on abdomen scans, and brain on head scans. This segmentation mask is then used for subsequent extraction of 91 radiomic features[3] using PyRadiomics.[70] Note, that because the same segmentation (on high-dose scans) is used for all algorithms, shape-based features (e.g., voxel volume) were excluded for computation of the RFS.

[3] This includes features from the following classes (# of features): first order statistics (18), gray level co-occurrence matrix (24), gray level run length matrix (16), gray level size zone matrix (16), neighboring gray tone difference matrix (4), and gray level dependence matrix (13).

Furthermore, we evaluate the algorithms in terms of their ability to reconstruct lesions and using classical image quality metrics for CT (Sections 5.5 and 5.6).

4.5 | LDCT-hard benchmark dataset

In our experiments we find that the performance of all algorithms varies greatly, both between different exam types and images of the same exam type. The latter observation motivates us to derive a novel collection of test datasets, each of which being a subset of the *Low-dose CT and Projection Dataset*.[39] We refer to LDCT-hard-q%, as the subset containing the q% slices with lowest average SSIM across all evaluated methods. To not underrepresent anatomies for which methods achieve generally higher SSIMs (e.g., head), this subset is collected for each exam type separately.

5 | RESULTS

5.1 | Hyperparameter optimization

We first verify that all methods converged within the 50 iterations of Bayesian hyperparameter optimization (Figure 2). To this end, we evaluate for each method and iteration i the relative deviation RelDev_i from the best setting w.r.t. the SSIM on the validation set

$$\mathrm{RelDev}_i = 1 - \frac{\max_{j \leq i} \mathrm{SSIM}_j}{\max_j \mathrm{SSIM}_j} . \tag{7}$$

We find that hyperparameter optimization for most of the methods converged within the first 40 iterations and none of the methods improved in the last five iterations (cf. intercept with x-axis in Figure 2). For all methods $\mathrm{RelDev}_{i \geq 30} < 0.5\%$.

5.2 | Evaluation using standard image quality metrics

We then evaluate all algorithms using the following image quality metrics: SSIM, PSNR, and VIF (Table 2). For each method, we test if it performs significantly better or worse than the previously published best method, using the nonparametric *Mann-Whitney U* test[71] with significance level $\alpha = 5\%$. While we find that ResNet significantly outperforms previous methods on the chest data, none of the newer methods consistently outperforms RED-CNN, one of the earliest deep learning-based methods for LDCT denoising (cf. **bold** numbers in Table 2). On the contrary, for many configurations newer methods perform significantly worse than RED-CNN (cf. *italic* numbers in Table 2). In particular, we find that the

TABLE 2 Quantitative evaluation using the metrics SSIM, PSNR, and VIF.

	Chest (10% dose)			Abdomen (25% dose)			Head (25% dose)			
	SSIM	**PSNR (dB)**	**VIF**	**SSIM**	**PSNR (dB)**	**VIF**	**SSIM**	**PSNR (dB)**	**VIF**	**Rank**
LD	0.34	18.77	0.09	0.84	28.67	0.34	0.88	26.4	0.55	9
CNN-10	**0.5867 ± 0.0006**	**27.71 ± 0.02**	**0.1915 ± 0.0008**	**0.896 ± 0.001**	**32.4 ± 0.1**	**0.449 ± 0.003**	**0.896 ± 0.004**	**28.9 ± 0.6**	**0.620 ± 0.006**	3
RED-CNN	**0.609 ± 0.002**	**28.36 ± 0.03**	**0.221 ± 0.003**	**0.9028 ± 0.0007**	**33.22 ± 0.07**	**0.491 ± 0.008**	**0.904 ± 0.001**	**30.4 ± 0.2**	**0.69 ± 0.01**	1
WGAN-VGG	*0.51 ± 0.03*	*25.5 ± 0.2*	*0.148 ± 0.004*	*0.882 ± 0.002*	*30.5 ± 0.9*	*0.38 ± 0.01*	*0.88 ± 0.02*	*25 ± 3*	*0.53 ± 0.02*	6†
ResNet	**0.610 ± 0.001**	**28.42 ± 0.03**	**0.224 ± 0.002**	*0.901 ± 0.002*	*33.15 ± 0.08*	*0.487 ± 0.006*	*0.901 ± 0.005*	*29.6 ± 0.8*	*0.67 ± 0.02*	2
QAE	*0.584 ± 0.003*	*27.62 ± 0.09*	*0.186 ± 0.003*	*0.894 ± 0.002*	*32.0 ± 0.2*	*0.418 ± 0.007*	*0.899 ± 0.001*	*28.5 ± 0.3*	*0.594 ± 0.008*	5
DU-GAN	*0.565 ± 0.004*	*26.7 ± 0.1*	*0.168 ± 0.002*	*0.894 ± 0.002*	*32.1 ± 0.3*	*0.427 ± 0.005*	0.903 ± 0.003	*29 ± 1*	*0.622 ± 0.005*	4
TransCT	*0.563 ± 0.002*	*26.99 ± 0.05*	*0.167 ± 0.002*	*0.877 ± 0.003*	*30.5 ± 0.2*	*0.372 ± 0.007*	*0.849 ± 0.005*	*24.7 ± 0.4*	*0.44 ± 0.01*	6†
Bilateral	*0.555 ± 0.001*	*25.59 ± 0.04*	*0.159 ± 0.002*	*0.859 ± 0.003*	*27.1 ± 0.1*	*0.361 ± 0.003*	*0.873 ± 0.002*	*26.6 ± 0.1*	*0.500 ± 0.004*	8

Note: We highlighted a metric in **bold**, if it is significantly better than the previously published best method on that anatomy. Likewise, we highlighted a metric in *italics*, if it is significantly worse than the previously published best method on that anatomy. The rank column (last column) is the competition ranking over all anatomies and metrics. We indicate a tie with †.

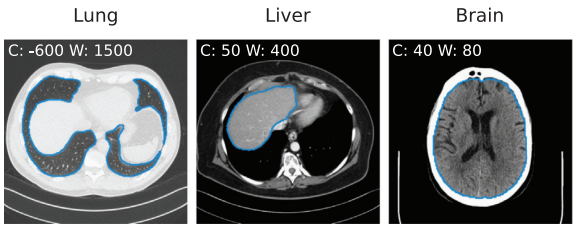


FIGURE 3 Contour plots of automatic segmentations for three high-dose scans of the test set of lung, liver, and brain. Radiomic features were extracted within these segmentations for low- and high-dose as well as all denoised volumes.

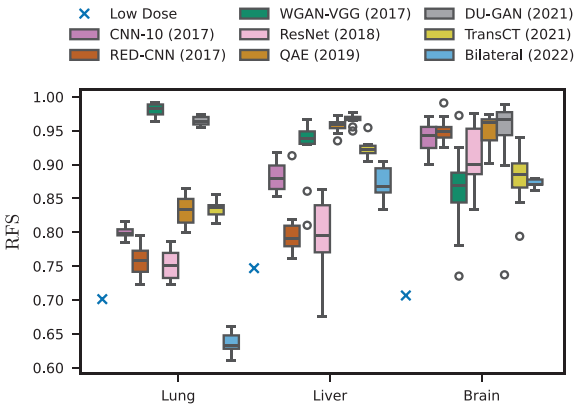


FIGURE 4 Radiomic feature similarity for different exam types and methods. Individual samples correspond to mean RFS over all scans of an anatomy for a single trained network. Box plots were then drawn over the 10 training runs with different random seeds (cf. Section 4.3.).

two newest methods considered in this study (TransCT and Bilateral) perform significantly worse w.r.t. all metrics and exam types compared to RED-CNN. Remarkably, they even perform significantly worse compared to the low-dose scan on few metric and exam type combinations (e.g., TransCT on head scans for all metrics; Bilateral on abdomen scans for PSNR).

5.3 | Evaluation using radiomic feature similarity

We further evaluate all algorithms using the radiomic feature similarity in order to better assess whether the differences observed in the previous section translate to clinical features.

In Figure 3 we show contour plots of the automatic segmentations of the brain, lung, and liver for three high-dose scans of the test set. We visually verify that segmentations are reasonably good for all 15 scans in the test set. Those segmentation masks are then used to extract radiomic features for all low- and high-dose, as well as all denoised volumes of the test set. Using the same segmentation mask for subsequent radiomic feature extraction of all algorithms ensures a fair comparison, despite possible small errors produced by the automatic segmentation pipeline.

Upon evaluation of the radiomic feature similarity (Table 3 and Figure 4), we find that radiomic features extracted for all denoising methods are significantly more similar to those extracted on the high-dose scan, compared to features extracted on the low-dose scan, with Bilateral on lung data being the only exception. We also find that contrary to our findings using standard image quality metrics, RED-CNN is outperformed by numerous other algorithms, including the (older) CNN-10, and newer algorithms such as WGAN-VGG and QAE. Remarkably, the two GAN-based algorithms WGAN-VGG and DUGAN outperform all other algorithms on the lung data by a large margin. We hypothesize that this is due to the lower dose (10% vs. 25% for all other anatomies) on that data and the ability of GANs to produce more realistic noise textures in high-ambiguity settings compared to methods trained with standard pixelwise loss functions.[72] Nonetheless, we do not find newer algorithms to consistently outperform older ones, and particularly the two newest algorithms considered

TABLE 3 Quantitative evaluation using the radiomic feature similarity.

	Lung	Liver	Brain	Rank
LD	0.7	0.75	0.71	9
CNN-10 (2017)	**0.800 ± 0.009**	**0.88 ± 0.02**	**0.94 ± 0.02**	4†
RED-CNN (2017)	*0.76 ± 0.02*	*0.80 ± 0.04*	0.95 ± 0.02	6
WGAN-VGG (2017)	**0.98 ± 0.01**	**0.92 ± 0.05**	*0.86 ± 0.07*	4†
ResNet (2018)	*0.75 ± 0.02*	*0.79 ± 0.06*	0.91 ± 0.05	7
QAE (2019)	*0.83 ± 0.02*	**0.96 ± 0.01**	0.95 ± 0.02	2
DU-GAN (2021)	*0.965 ± 0.007*	**0.967 ± 0.008**	0.94 ± 0.08	1
TransCT (2021)	*0.83 ± 0.01*	*0.92 ± 0.01*	*0.88 ± 0.04*	3
Bilateral (2022)	*0.64 ± 0.01*	*0.87 ± 0.02*	*0.873 ± 0.006*	8

Note: **Bold** numbers indicate that a method is significantly better than the previously published best method on that anatomy. Likewise, italics indicate that it is significantly worse. The rank column (last column) is the competition ranking over all anatomies and metrics. We indicate a tie with †.

in our study (TransCT and Bilateral) perform significantly worse w.r.t. radiomic feature similarity of all organs compared to older methods. Figure 5 shows qualitative results for the slices from the test dataset for which the average SSIM over all methods is lowest (−) and highest (+), respectively. As can be seen, for each anatomy, the slice maximizing the average SSIM is the one where the cross sectional area of the patient is small, thus reducing the noise in the low-dose image.

5.4 | Evaluation on LDCT-hard datasets

Figure 6 shows the performance of individual methods for increasingly hard subsets of the training data (i.e., smaller q). We find a strong correlation between metrics for each method and the low-dose scan. Although not surprising, this indicates that methods perform increasingly worse for increasing deviations of the low-dose scan from the high-dose scan. Additionally, the ranking among methods remains mostly invariant to q, and thus we conclude that all methods are similar in terms of their robustness to different amounts of deterioration of the low-dose scan. Remarkably, WGAN-VGG, having a lower VIF and PSNR compared to the low-dose scan on head exams for the regular test set (corresponding to $q = 100\%$), has a higher VIF and PSNR compared to the low-dose scan for more difficult slices ($q \leq 16\%$ for VIF, $q \leq 40\%$ for PSNR). This may be explained by the aforementioned ability of GANs to produce more realistic results in high-ambiguity settings compared to networks trained in a pixelwise fashion.

5.5 | Evaluation on lesions

The main downstream task for clinical low-dose CT is the detection and diagnosis of lesions. To better assess whether the denoising algorithms improve performance on this downstream task compared to low-dose reconstructions we utilize the lesion annotations provided with the *LDCT Image and Projection dataset*.[39] For our test set there exist a total of eleven annotations covering all three exam types and six different diagnosis. For each of these lesions we compute the RMSE and PSNR compared to the high-dose reconstruction within the bounding-box surrounding the lesion (Table 4).

Here, we find that all methods have lower deviations from the ground truth compared to the low-dose reconstruction. We also find that the ranking mostly agrees with the ranking based on standard image quality metrics (cf. Table 2) and in particular w.r.t. the three best performing methods (RED-CNN > ResNet > CNN-10) both rankings agree. We provide reconstruction results for all lesions and algorithms in the Appendix, Figures C.4 to C.6.

5.6 | Evaluation using physical CT IQA metrics

We also analyzed the algorithms using physical image quality metrics, a common way to evaluate the technical performance of CT systems. A discussion on the limitations of these metrics is provided in Section 2.2. In particular, we perform all evaluations using patient scans of the test set instead of phantom measurements to avoid an out-of-distribution setting. In the following, we provide the main results of this evaluation and refer to Appendix D for additional results.

Contrast-to-noise ratio for liver lesion:
To evaluate the algorithms' capability to recover low-contrast structures, we compute the CNR for one liver lesion of the test set (cf. Section 5.5 and Figure C.5, lesion #5). To this end, we place one circular region of interest (ROI) in the lesion and one in the surrounding homogeneous liver tissue. Here we find that all methods improve the CNR compared to the low-dose reconstruction (Table 5). Remarkably, most methods improve the

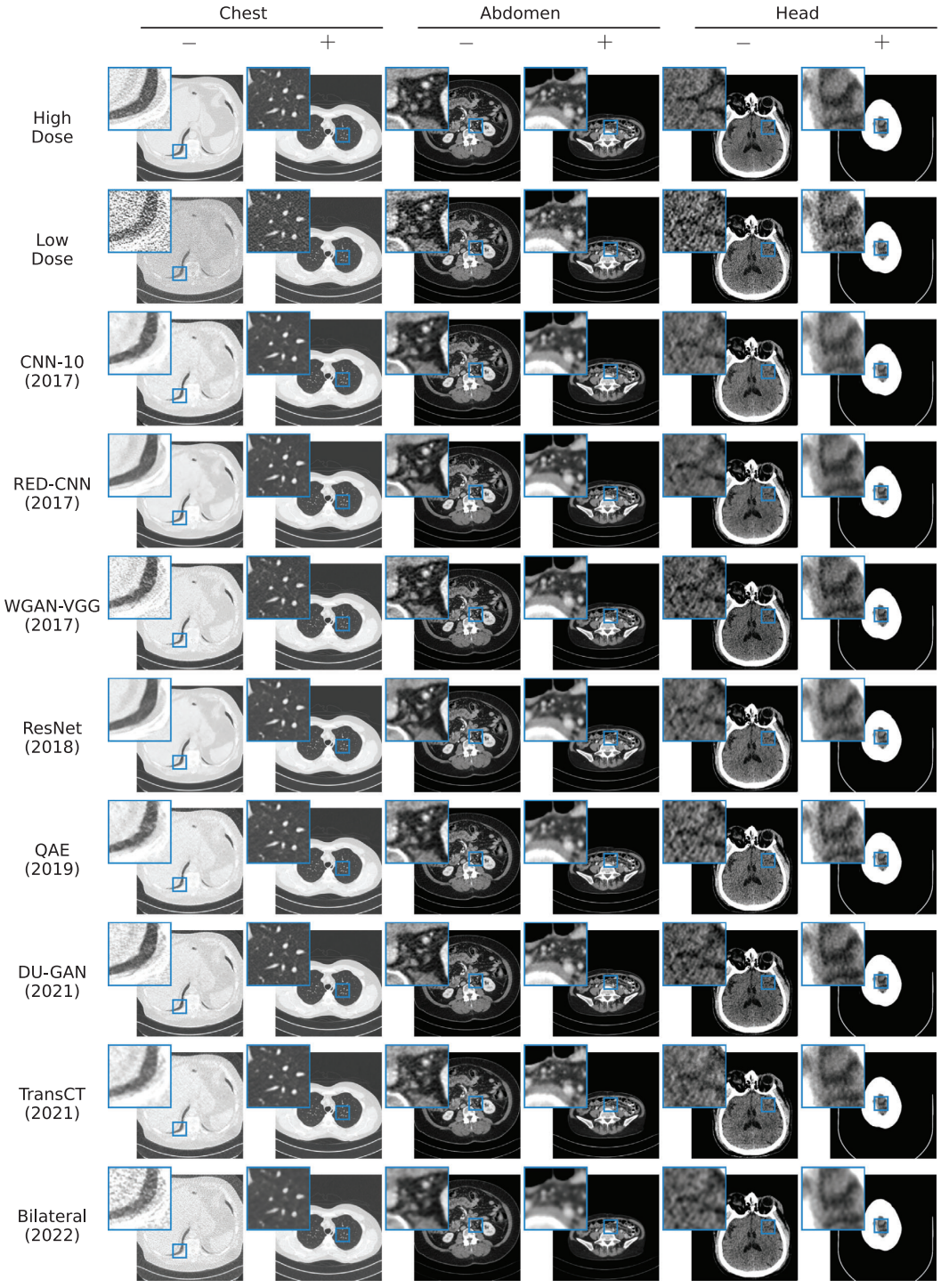


FIGURE 5 Best viewed zoomed in. Slices from the test dataset, for which the average SSIM over all methods is lowest (−) and highest (+). For each method, we show results for the best performing network (over the 10 random trials), that is, network having the highest SSIM on the validation data.

TABLE 4 Quantitative evaluation of the algorithms ability to reconstruct lesions for the three anatomies, averaged over lesions.

	Chest (10% dose)		Abdomen (25% dose)		Head (25% dose)		
	PSNR (dB)	RMSE (HU)	PSNR (dB)	RMSE (HU)	PSNR (dB)	RMSE (HU)	Rank
LD	9.67	169.21	12.39	23.68	9.76	8.43	9
CNN-10 (2017)	13.2	73.24	14.59	14.23	11.89	5.15	3
RED-CNN (2017)	**13.36**	**71.98**	**14.87**	**13.39**	**12.28**	**4.74**	1
WGAN-VGG (2017)	12.11	94.79	14.02	16.41	10.65	6.85	8
ResNet (2018)	13.28	73.15	14.82	13.53	12.1	4.92	2
QAE (2019)	12.94	77.93	14.2	15.54	11.28	5.93	6†
DU-GAN (2021)	13.04	76.27	14.46	14.73	11.06	6.23	5
TransCT (2021)	13.04	82.57	14.41	14.94	11.12	6.12	6†
Bilateral (2022)	12.6	84.53	14.72	13.92	12.05	4.98	4

Note: We indicate the best performing method for an anatomy and metric in **bold**. The rank column (last column) is the competition ranking over all anatomies and metrics. We indicate a tie with†.

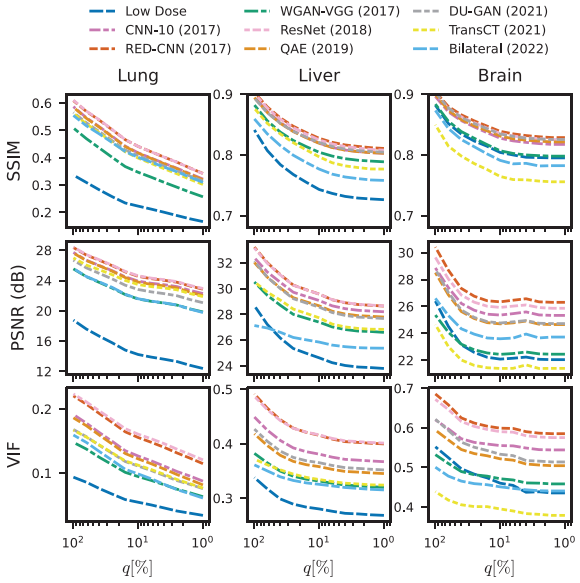


FIGURE 6 Evaluation of all methods for LDCT-hard-q% for different values of q (right is smaller). For some settings and anatomies, methods perform up to 50% worse for small q. The regular test set corresponds to $q = 100\%$. Errorbars were omitted to improve visibility.

CNR even compared to the high-dose reconstruction which is likely due to the pixelwise loss, which smooths the image and thus reduces noise.

CT number accuracies:

We also evaluate the algorithms' ability to recover the CT numbers of the high-dose reconstruction. To this end, we place five ROIs each for three of the chest exams in muscle tissue and compute the mean CT number for each reconstruction and ROI. We then compute the absolute deviation from the mean CT number of the high-dose reconstruction and show the results in Figure 7. The mean CT number over all ROIs of the high-dose scans is 49.76 ± 6.30 HU. We find that three of the eight algorithms (RED-CNN, ResNet, and Bilateral) perform significantly better than the low-dose reconstruction in recovering the CT numbers of the

TABLE 5 Quantitative evaluation of the CNR for one liver metastasis (#5 in Figure C.5).

	CNR	Ranking
HD	2.17	5
LD	0.85	9
CNN-10 (2017)	2.47	4
RED-CNN (2017)	**3.13**	1
WGAN-VGG (2017)	1.52	8
ResNet (2018)	2.93	2
QAE (2019)	1.63	7
DU-GAN (2021)	1.81	6
TransCT (2021)	2.24	5
Bilateral (2022)	2.74	3

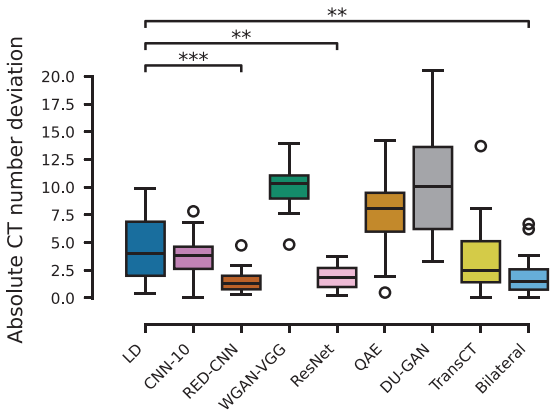


FIGURE 7 CT number accuracy over 15 ROIs in muscle tissue of chest exams. Statistical significance is indicated with $*$: $p < 0.05$, $**$: $p < 0.01$, $***$: $p < 0.001$.

high-dose reconstruction with RED-CNN achieving the lowest mean deviation with 1.58 ± 1.19 HU. Significance was tested using a Wilcoxon signed-rank test. The two GAN-based methods (WGAN-VGG and DU-GAN) perform worst in this regard, which can be attributed to the fact that they are not trained exclusively using a pixelwise loss and adversarial losses do not directly enforce gray value consistency.

Line profile analysis:
The spatial resolution of an imaging system is commonly evaluated using the MTF. However, since the MTF is not well-defined for nonlinear algorithms and the task transfer function (TTF) requires phantom measurements, we perform an assessment of the algorithms' ability to recover sharp edges in the image using line profiles. Here we find that while all algorithms reduce the noise compared to the LD reconstruction and stay closer to the high-dose reconstruction, some algorithms fail to recover sharp edges in the line profile. We provide the line profiles and additional details in Figure D.7 and Appendix D.

6 | DISCUSSION

In this study, we revisited some of the numerous proposed deep learning-based algorithms for low-dose CT image denoising. We discovered several limitations in the experimental setups of these methods that hinder the verifiability of their claimed improvements. To overcome these challenges, we proposed a novel benchmark setup that promotes fair and reproducible evaluations. The setup comprises a unified data preprocessing, rigorous hyperparameter optimization, and evaluation using various metrics, including a novel metric that measures the similarity of radiomic features between the denoised volume and the high-dose scan.

Upon evaluation of eight deep learning-based denoising algorithms proposed over the past six years, we find that there has been little progress. Particularly, when evaluated using standard image quality measures such as SSIM and PSNR, we find that no method consistently outperforms one of the earliest methods, RED-CNN. When evaluated using the radiomic feature similarity, we find that algorithms trained with an adversarial loss significantly outperform methods trained with pixelwise losses on some data, indicating that the radiomic feature similarity provides useful information beyond standard, nonclinical image quality metrics. Nonetheless, the newest algorithms considered in our study fail to consistently outperform older ones. An evaluation on lesion annotations and using physical image quality assessment metrics leads to the same conclusion. We also evaluated all methods on subsets of the test data consisting of increasingly difficult slices and find that methods are similarly robust to different amounts of deterioration of the low-dose scan.

We note that our evaluation mainly focused on distortion (full-reference) measures[73] and that the hyperparameter optimization is limited to a single such distortion measure, the SSIM. Future work should consider including more perceptual measures (e.g., based on feature maps of DNNs) both for hyperparameter optimization and subsequent evaluation of the algorithms. This is particularly important, given the recent shift towards using more perceptual loss functions in the field. Other possible extensions include evaluation of more algorithms including score-based methods[74,75] and methods that leverage multiple axial slices[74–77] as well as training and/or evaluation on more datasets, particularly those that contain lesion annotations.

Similar to "reality checks" in related fields,[78,79] our study highlights the need for a more rigorous and fair evaluation of novel deep learning-based denoising methods for low-dose CT image denoising. We believe that our benchmark setup is a first and important step towards this direction and will help to develop novel and better algorithms.

ACKNOWLEDGMENTS
This work was supported in part by the Helmholtz International Graduate School for Cancer Research, Heidelberg, Germany.

CONFLICT OF INTEREST STATEMENT
The authors declare no conflicts of interest.

REFERENCES
1. Kalra MK, Maher MM, Toth TL, et al. Strategies for CT radiation dose optimization. *Radiology*. 2004;230:619-628.
2. Brenner DJ, Hall EJ. Computed tomography–an increasing source of radiation exposure. *N Engl J Med*. 2007;357:2277-2284.
3. Ziegler A, Koehler T, Proksa R. Noise and resolution in images reconstructed with FBP and OSC algorithms for CT. *Med Phys*. 2007;34:585-598.
4. Sidky EY, Kao C, Pan X. Accurate image reconstruction from few–views and limited–angle data in divergent–beam CT. *J X Ray Sci Technol*. 2006;14:119-139.
5. Balda M, Hornegger J, Heismann B. Ray contribution masks for structure adaptive sinogram filtering. *IEEE Trans Med Imaging*. 2012;31:1228-1239.
6. Feruglio PF, Vinegoni C, Gros J, Sbarbati A, Weissleder R. Block matching 3D random noise filtering for absorption optical projection tomography. *Phys Med Biol*. 2010;55:5401-5415.
7. Li Z, Yu L, Trzasko JD, Lake DS, Blezek DJ, Fletcher JG, McCollough CH, Manduca A. Adaptive nonlocal means filtering based on local noise level for CT denoising. *Med Phys*. 2014;41:011908.
8. Manduca A, Yu L, Trzasko JD, et al. Projection space denoising with bilateral filtering and CT noise modeling for dose reduction in CT. *Med Phys*. 2009;36:4911-4919.
9. Sukovic P, Clinthorne N. Penalized weighted least-squares image reconstruction for dual energy x-ray transmission tomography. *IEEE Trans Med Imaging*. 2000;19:1075-1081.

10. Chen H, Zhang Y, Zhang W, Liao P, Li K, Zhou J, Wang G. Low-dose CT denoising with convolutional neural network. In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE; 2017:143-146.
11. Chen H, Zhang Y, Zhang W, et al. Low-dose CT via convolutional neural network. *Biomed Opt Express*. 2017;8:679-694.
12. Chen H, Zhang Y, Kalra MK, et al. Low-dose CT with a residual encoder-decoder convolutional neural network. *IEEE Trans Med Imaging*. 2017;36:2524-2535.
13. Fan F, Shan H, Kalra MK, et al. Quadratic autoencoder (Q-AE) for low-dose CT denoising. *IEEE Trans Med Imaging*. 2020;39:2035-2050.
14. Heinrich MP, Stille M, Buzug TM. Residual U-Net convolutional neural network architecture for low-dose CT denoising. *Curr Dir Biomed Eng*. 2018;4:297-300.
15. Huang Z, Zhang J, Zhang Y, Shan H. DU-GAN: generative adversarial networks with dual-domain U-Net-based discriminators for low-dose CT denoising. *IEEE Trans Instrum Meas*. 2022;71:1-12.
16. Kang E, Min J, Ye JC. A deep convolutional neural network using directional wavelets for low-dose x-ray CT reconstruction. *Med Phys*. 2017;44:e360-e375.
17. Ramanathan S, Ramasundaram M. Low dose CT image reconstruction using deep convolutional residual learning network. *SN Comput Sci*. 2023;4:720.
18. Shan H, Padole A, Homayounieh F, et al. Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction. *Nat Mach Intell*. 2019;1:269-276.
19. Wagner F, Thies M, Gu M, et al. Ultralow-parameter denoising: trainable bilateral filter layers in computed tomography. *Med Phys*. 2022;49:5107-5120.
20. Yang Q, Yan P, Zhang Y, et al. Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. *IEEE Trans Med Imaging*. 2018;37:1348-1357.
21. Yang S, Pu Q, Lei C, Zhang Q, Jeon S, Yang X. Low-dose CT denoising with a high-level feature refinement and dynamic convolution network. *Med Phys*. 2023;50:3597-3611.
22. Zhang Z, Yu L, Liang X, Zhao W, Xing L. TransCT: dual-path transformer for low dose computed tomography. In: *International Conference on Medical Image Computing and Computer Assisted Intervention*. MICCAI; 2021.
23. Humphries T, Coulter S, Si D, Simms M, Xing R. Comparison of deep learning approaches to low dose CT using low intensity and sparse view data. In: Bosmans H, Chen G-H, Gilat Schmidt T, eds. *Medical Imaging 2019: Physics of Medical Imaging*. SPIE; 2019:156.
24. Ma Y-J, Ren Y, Feng P, He P, Guo X-D, Wei B. Sinogram denoising via attention residual dense convolutional neural network for low-dose computed tomography. *Nucl Sci Tech*. 2021;32:41.
25. Yang L, Li Z, Ge R, Zhao J, Si H, Zhang D. Low-dose CT denoising via sinogram inner-structure transformer. *IEEE Trans Med Imaging*. 2023;42:910-921.
26. Zainulina E, Chernyavskiy A, Dylov DV. No-reference denoising of low-dose CT projections. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE; 2021:77-81.
27. Hong Z, Zeng D, Tao X, Ma J. Learning CT projection denoising from adjacent views. *Med Phys*. 2023;50:1367-1377.
28. Cormack AM. Representation of a function by its line integrals, with some radiological applications. *J Appl Phys*. 1963;34:2722-2727.
29. Gordon R, Bender R, Herman GT. Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and x-ray photography. *J Theor Biol*. 1970;29:471-481.
30. Andersen AH, Kak AC. Simultaneous algebraic reconstruction technique (SART): a superior implementation of the art algorithm. *Ultrason Imaging*. 1984;6:81-94.
31. Missert AD, Leng S, Yu L, McCollough CH. Noise subtraction for low-dose CT images using a deep convolutional neural network. In: *Proceedings of the Fifth International Conference on Image Formation in X-Ray Computed Tomography, Salt Lake City, UT, USA*. 2018:399-402.
32. Baguer DO, Leuschner J, Schmidt M. Computed tomography reconstruction using deep image prior and learned reconstruction methods. *Inverse Probl*. 2020;36:094004.
33. Yin X, Zhao Q, Liu J, et al. Domain progressive 3D residual convolution network to improve low-dose CT imaging. *IEEE Trans Med Imaging*. 2019;38:2903-2913.
34. Chao L, Zhang P, Wang Y, Wang Z, Xu W, Li Q. Dual-domain attention-guided convolutional neural network for low-dose cone-beam computed tomography reconstruction. *Knowledge Based Syst*. 2022;251:109295.
35. Zhang Y, Hu D, Zhao Q, et al. CLEAR: comprehensive learning enabled adversarial reconstruction for subtle structure enhanced low-dose CT imaging. *IEEE Trans Med Imaging*. 2021;40:3089-3101.
36. Zhou B, Zhou SK, Duncan JS, Liu C. Limited view tomographic reconstruction using a cascaded residual dense spatial-channel attention network with projection data fidelity layer. *IEEE Trans Med Imaging*. 2021;40:1792-1804.
37. Zhou B, Chen X, Xie H, Zhou SK, Duncan JS, Liu C. DuDoUFNet: dual-domain under-to-fully-complete progressive restoration network for simultaneous metal artifact reduction and low-dose CT reconstruction. *IEEE Trans Med Imaging*. 2022;41:3587-3599.
38. McCollough CH, Bartley AC, Carter RE, et al. Low-dose CT for the detection and classification of metastatic liver lesions: results of the 2016 low dose CT grand challenge. *Med Phys*. 2017;44:e339-e352.
39. McCollough C, Chen B, Holmes III DR, et al. Low dose CT image and projection data (data set). The Cancer Imaging Archive. 2020. https://doi.org/10.7937/9NPB-2637
40. Divel SE, Pelc NJ. Accurate image domain noise insertion in CT images. *IEEE Trans Med Imaging*. 2020;39:1906-1916.
41. Horenko I, Pospíšil L, Vecchi E, et al. Low-cost probabilistic 3D denoising with applications for ultra-low-radiation computed tomography. *J Imaging*. 2022;8:156.
42. Wang Z, Bovik A, Sheikh H, Simoncelli E. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*. 2004;13:600-612.
43. Verdun FR, Racine D, Ott JG, et al. Image quality in CT: from physical measurements to model observers. *Physica Med*. 2015;31:823-843.
44. Renieblas GP, Nogués AT, González AM, Gómez-Leon N, Del Castillo EG. Structural similarity index family for image quality assessment in radiological images. *J Med Imaging*. 2017;4:035501.
45. Ohashi K, Nagatani Y, Yoshigoe M, et al. Applicability evaluation of full-reference image quality assessment methods for computed tomography images. *J Imaging Inform Med*. 2023;36:2623-2634.
46. Mason A, Rioux J, Clarke SE, et al. Comparison of objective image quality metrics to expert radiologists' scoring of diagnostic quality of MR images. *IEEE Trans Med Imaging*. 2020;39:1064-1072.
47. Sheikh H, Bovik A. Image information and visual quality. *IEEE Trans Image Process*. 2006;15:430-444.
48. Pan S, Flores J, Lin CT, Stayman JW, Gang GJ. Generative adversarial networks and radiomics supervision for lung lesion synthesis. *Proc SPIE-Int Soc Opt Eng*. 2021;11595:115950O.
49. Wei L, Hsu W. Efficient and accurate spatial-temporal denoising network for low-dose CT scans. In: *Medical Imaging with Deep Learning*. 2021.
50. Patwari M, Gutjahr R, Marcus R, et al. Reducing the risk of hallucinations with interpretable deep learning models for low-dose CT denoising: comparative performance analysis. *Phys Med Biol*. 2023;68:19LT01.

51. Barrett HH, Myers KJ. *Foundations of Image Science*. Wiley; 2004.
52. Richard S, Husarik DB, Yadava G, Murphy SN, Samei E. Towards task-based assessment of CT performance: system and object MTF across different reconstruction algorithms. *Med Phys*. 2012;39:4115-4122.
53. Vaishnav JY, Jung WC, Popescu LM, Zeng R, Myers KJ. Objective assessment of image quality and dose reduction in CT iterative reconstruction. *Med Phys*. 2014;41:071904.
54. Samei E, Bakalyar D, Boedeker KL, et al. Performance evaluation of computed tomography systems: summary of AAPM Task Group 233. *Med Phys*. 2019;46:e735-e756.
55. Hsieh J, Liu E, Nett B, Tang J, Thibault J-B, Sahney S. *A New Era of Image Reconstruction: TrueFidelity™*. White Paper. GE Healthcare; 2019.
56. Franzen R. Kodak Lossless True Color Image Suite (data set). 1999. https://r0k.us/graphics/kodak/
57. Martin D, Fowlkes C, Tal D, Malik J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: *Proceedings Eighth IEEE International Conference on Computer Vision (ICCV 2001)*. Vol 2. IEEE; 2001:416-423.
58. Zhang L, Wu X, Buades A, Li X. Color demosaicking by local directional interpolation and nonlocal adaptive thresholding. *J Electron Imaging*. 2011;20:023016.
59. Huang J-B, Singh A, Ahuja N. Single image super-resolution from transformed self-exemplars. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2015:5197-5206.
60. Zhang K, Zuo W, Chen Y, Meng D, Zhang L. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. *IEEE Trans Image Process*. 2017;26:3142-3155.
61. Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 24. Curran Associates, Inc.; 2011.
62. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *J Mach Learn Res*. 2012;13:281-305.
63. Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc.; 2012.
64. Liu C, Gao C, Xia X, Lo D, Grundy J, Yang X. On the reproducibility and replicability of deep learning in software engineering. *ACM Trans Softw Eng Methodol*. 2021;31:15:1-15:46.
65. Kc P, Zeng R, Farhangi MM, Myers KJ. Deep neural networks-based denoising models for CT imaging and their efficacy. In: *Medical Imaging 2021: Physics of Medical Imaging*. Vol. 11595. SPIE; 2021:105-117.
66. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: *Advances in Neural Information Processing Systems*. Vol. 27. Curran Associates, Inc.; 2014.
67. Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: *International Conference on Machine Learning (ICML)*. PMLR; 2017:214-223.
68. Schonfeld E, Schiele B, Khoreva A. A U-Net based discriminator for generative adversarial networks. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2020:8204-8213.
69. Wasserthal J, Breit H-C, Meyer MT, et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiol Artif Intell*. 2023;5:e230024.
70. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77:e104-e107.
71. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat*. 1947;18:50-60.
72. Zhao H, Gallo O, Frosio I, Kautz J. Loss functions for image restoration with neural networks. *IEEE Trans Comput Imaging*. 2017;3:47-57.
73. Blau Y, Michaeli T. The perception-distortion tradeoff. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE; 2018:6228-6237.
74. Li Q, Li C, Yan C, et al. Ultra-low dose CT image denoising based on conditional denoising diffusion probabilistic model. In: *2022 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*. 2022:198-205.
75. Gao Q, Li Z, Zhang J, Zhang Y, Shan H. CoreDiff: contextual error-modulated generalized diffusion model for low-dose CT denoising and generalization. *IEEE Trans Med Imaging*. 2024;43:745-759.
76. Zhou Z, Huber NR, Inoue A, McCollough CH, Yu L. Multislice input for 2D and 3D residual convolutional neural network noise reduction in CT. *J Med Imaging*. 2023;10:014003.
77. Shi J, Elkilany O, Fischer A, Suppes A, Pelt DM, Batenburg KJ. Lodoind: introducing a benchmark low-dose industrial CT dataset and enhancing denoising with 2.5D deep learning techniques. In: *13th Conference on Industrial Computed Tomography (iCT), Wels Campus, Austria*, 2024. https://doi.org/10.58286/29228
78. Melis G, Dyer C, Blunsom P. On the state of the art of evaluation in neural language models. In: *International Conference on Learning Representations*. 2018.
79. Musgrave K, Belongie S, Lim S-N. A metric learning reality check. In: Vedaldi A, Bischof H, Brox T, Frahm J-M, eds. *Computer Vision – ECCV 2020*. Lecture Notes in Computer Science. Springer International Publishing; 2020:681-699.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Eulig E, Ommer B, Kachelrieß M. Benchmarking deep learning-based low-dose CT image denoising algorithms. *Med Phys*. 2024;1-13. https://doi.org/10.1002/mp.17379

Chapter 3

Invariances of Low-Dose CT Denoising Networks

E. Eulig, F. Jäger, J. Maier, B. Ommer, and M. Kachelrieß. "Reconstructing and Analyzing the Invariances of Low-Dose CT Image Denoising Networks". In: *Medical Physics* 52.1 (2025), pp. 188–200

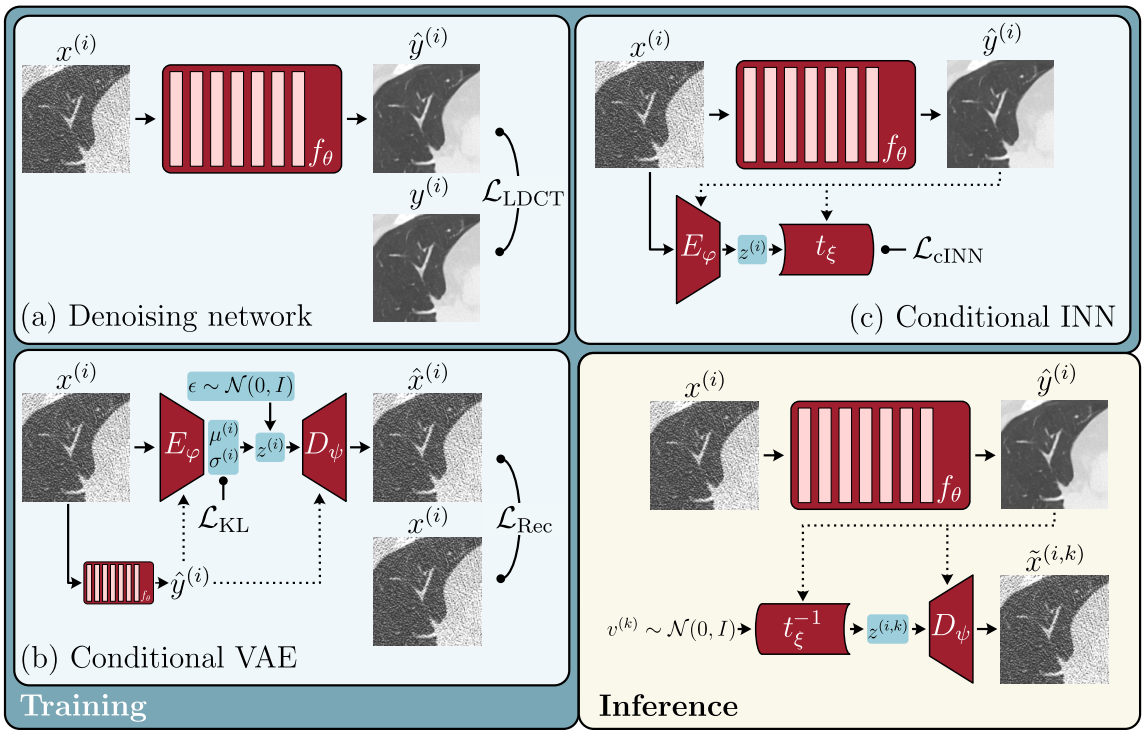
This chapter presents our work on reconstructing and analyzing the invariances of LDCT image denoising networks. We first summarize it in Sec. 3.1 and provide the full paper [69] in the subsequent Sec. 3.2.

3.1 Summary

3.1.1 Introduction

Similar to the project discussed in the previous chapter (Chapter 2), this project is concerned with evaluating deep learning-based LDCT image denoising algorithms. However, instead of focusing on the performance (in terms of quantitative image measures) of these algorithms, we here aim to analyze the invariances that these algorithms have. Invariances are a crucial property of any machine learning (ML) model and are defined as transformations of the input data that do not change the prediction of the model.

In most cases, such invariances are a desired property of a model, allowing them to model non-injective functions. E.g., for most applications, an object classifier should be invariant to translations of this object within the input image as this does not change the object's identity. For the task of LDCT image denoising, such desired invariances would be the invariance to noise, or streak artifacts in the image. Besides desired invariances, models can also have undesired invariances, which can lead to biases in the model predictions. An undesired invariance of a LDCT image denoising network could be the invariance to specific anatomical structures, which could lead to the removal or hallucination of diagnostic features in the image.



▲ **Figure 3.1.** Pipeline to reconstruct invariances of LDCT image denoising networks. We first train (a) the denoising network f_θ to map low-dose images $x^{(i)}$ to high-dose images $y^{(i)}$, (b) a variational autoencoder (VAE) to learn a complete representation of a low-dose image $x^{(i)}$, conditional on the denoised image $\hat{y}^{(i)}$, and (c) a conditional invertible neural network (cINN) to disentangle the invariances of the denoising network f_θ. To generate invariance images we sample from the invariance distribution $p(v)$ and first apply the inverse of the cINN and then the decoder of the VAE. Solid arrows represent inputs/outputs to modules, dotted arrows represent conditional inputs to a module. We indicate points where loss functions $\mathcal{L}$ are calculated using $\bullet\!\!-\mathcal{L}$. Figure taken from [69].

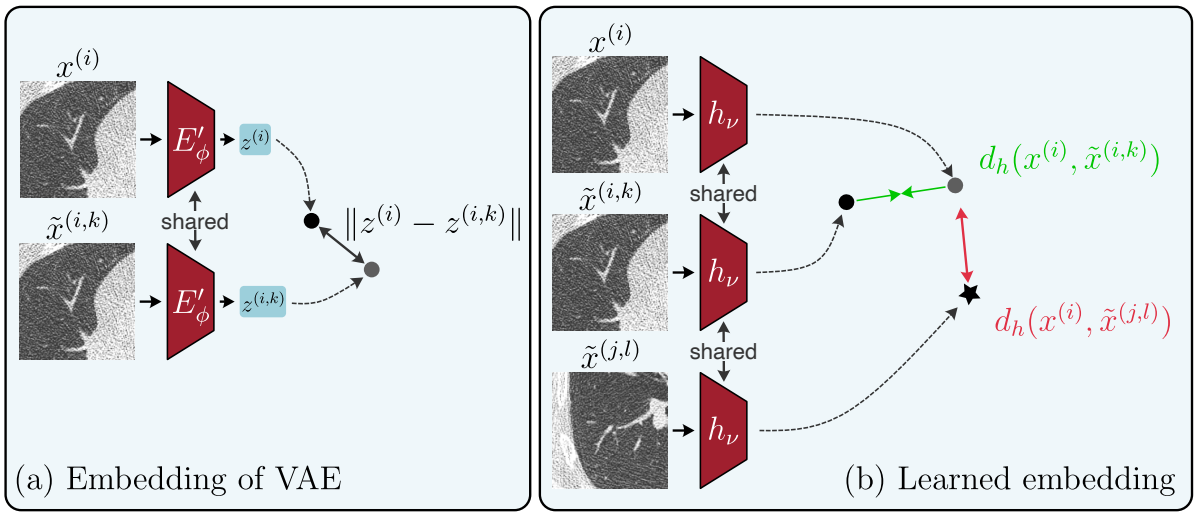
Identifying and analyzing the invariances of a model can provide insights into the model's behavior and can help to identify potential biases or limitations of the model or the training data. This is important to improve the interpretability and robustness of deep learning-based methods for medical imaging in general and LDCT image denoising, a prevalent application of DNNs for medical imaging, in particular.

Our method to reconstruct the invariances is based on previous work by Rombach et al. [70], who reconstructed the invariances of image classification networks. In this work, we adapted their method to the task of LDCT image denoising, developed suitable methods to analyze the reconstructed invariances, and used the resulting framework to evaluate four popular LDCT image denoising networks in terms of their invariances.

3.1.2 Methods

Our pipeline to reconstruct invariances of LDCT image denoising networks comprises a training and inference phase. An overview is provided in Fig. 3.1.

In the training phase, we first train some DNN f_θ, with θ being its parameters, to denoise



▲ **Figure 3.2.** Proposed methods to analyze sampled invariances. (a) based on the embedding of an unconditional VAE, and (b) based on an embedding learned via deep metric learning. Figure from [69].

LDCT images $x^{(i)}$ (Fig. 3.1, (a)). Typically, these networks are trained by minimizing some pixelwise or adversarial loss $\mathcal{L}_{\text{LDCT}}$ between network predictions $\hat{y}^{(i)} = f_\theta\left(x^{(i)}\right)$ and ground truth high-dose images $y^{(i)}$. We then use the predictions from this network as conditional input to a VAE $D_\psi \circ E_\varphi$ that is trained to reconstruct low-dose images (Fig. 3.1, (b)). As any VAE, this network is trained by minimizing a combination of a reconstruction loss $\mathcal{L}_{\text{Rec}}$ in image space and a Kullback-Leibler (KL) divergence $\mathcal{L}_{\text{KL}}$ in latent space $z^{(i)} = E_\varphi(x^{(i)}|\hat{y}^{(i)})$. The conditional input from the denoising network allows the VAE to focus on information in the input image that are not contained in the denoised image $\hat{y}^{(i)}$ already. In our experiments, we find that this improves the data representation compared to an unconditional VAE [71]. Finally, we train a cINN t_ξ to disentangle the information about the input image $x^{(i)}$, in the latent space $z^{(i)}$, to which f_θ is invariant to from the one it is not invariant to by learning a mapping $t_\xi : p(z^{(i)}|\hat{y}^{(i)}) \to p(v)$ with $p(v)$ being the distribution of invariances.

To then reconstruct invariances of f_θ for a given low-dose image $x^{(i)}$, we first denoise the image using f_θ to obtain $\hat{y}^{(i)}$. We then sample an invariance realization $v^{(k)} \sim p(v)$ and apply the inverse of the cINN t_ξ^{-1} to obtain a latent representation $z^{(i,k)}$ given $\hat{y}^{(i)}$. Applying the pretrained decoder D_ψ to this latent representation then yields the reconstructed invariance image $\tilde{x}^{(i,k)}$ (Fig. 3.1, bottom right). We can repeat this process for multiple samples $v^{(k)}$ to obtain a distribution of invariance images, i.e., images that only differ in the invariance realization.

To better assess whether the networks are invariant to anatomical structures or other image *content*, we further developed two methods to analyze sampled invariances (Fig. 3.2). The first method measures the similarity between input images $x^{(i)}$ and sampled invariances $\tilde{x}^{(i,k)}$ in the latent space of a separate unconditional VAE (Fig. 3.2, (a)). The second method uses a deep metric learning (DML) approach to learn a metric space in which an invariance image $\tilde{x}^{(i,k)}$ is closer to the input image $x^{(i)}$ than to invariance images $\tilde{x}^{(j,l)}, j \neq i$ of other input images $x^{(j)}$ (Fig. 3.2, (b)).

3.1.3 Results

We evaluated four popular LDCT image denoising networks using our proposed framework by generating and analyzing their invariances for random crops of a held-out test set. Here, we found that all networks are predominantly invariant to the noise level and noise realization of the input images. As discussed above, this is an expected and desired property of any denoising algorithm.

Evaluating whether the networks are invariant to anatomical structures or other content directly in the image space is difficult as differences in content can easily be overshadowed by differences in noise. We therefore used the two proposed similarity measures to find samples for which networks show content-related invariances. Using these, we found that indeed the networks are invariant to anatomical structures to some extent and that the two proposed similarity measures are suitable to quantify the amount of content-related invariances. We also performed a quantitative analysis of the invariances and found that most networks, despite showing significantly different denoising capabilities, have similar amounts of content-related invariances.

3.2 Paper: Reconstructing and Analyzing the Invariances of Low-Dose CT Image Denoising Networks

The following pages contain the full paper [69]. To improve readability, the supplementary material of this paper is provided in Appendix B.

DOI: 10.1002/mp.17413

RESEARCH ARTICLE

MEDICAL PHYSICS

Reconstructing and analyzing the invariances of low-dose CT image denoising networks

Elias Eulig[1,2] | **Fabian Jäger**[1,2] | **Joscha Maier**[1] | **Björn Ommer**[3] | **Marc Kachelrieß**[1,4]

[1]Division of X-Ray Imaging and Computed Tomography, German Cancer Research Center (DKFZ), Heidelberg, Germany

[2]Faculty of Physics and Astronomy, Heidelberg University, Heidelberg, Germany

[3]CompVis @ LMU Munich and MCML, Munich, Germany

[4]Medical Faculty Heidelberg, Heidelberg University, Heidelberg, Germany

Correspondence
Elias Eulig, Division of X-Ray Imaging and Computed Tomography, German Cancer Research Center (DKFZ), Heidelberg, Germany.
Email: elias.eulig@dkfz.de

Funding information
Helmholtz International Graduate School for Cancer Research

Abstract

Background: Deep learning-based methods led to significant advancements in many areas of medical imaging, most of which are concerned with the reduction of artifacts caused by motion, scatter, or noise. However, with most neural networks being black boxes, they remain notoriously difficult to interpret, hindering their clinical implementation. In particular, it has been shown that networks exhibit invariances w.r.t. input features, that is, they learn to ignore certain information in the input data.

Purpose: To improve the interpretability of deep learning-based low-dose CT image denoising networks.

Methods: We learn a complete data representation of low-dose input images using a conditional variational autoencoder (cVAE). In this representation, invariances of any given denoising network are then disentangled from the information it is not invariant to using a conditional invertible neural network (cINN). At test time, image-space invariances are generated by applying the inverse of the cINN and subsequent decoding using the cVAE. We propose two methods to analyze sampled invariances and to find those that correspond to alterations of anatomical structures.

Results: The proposed method is applied to four popular deep learning-based low-dose CT image denoising networks. We find that the networks are not only invariant to noise amplitude and realizations, but also to anatomical structures.

Conclusions: The proposed method is capable of reconstructing and analyzing invariances of deep learning-based low-dose CT image denoising networks. This is an important step toward interpreting deep learning-based methods for medical imaging, which is essential for their clinical implementation.

KEYWORDS
computed tomography, deep learning, explainability, invariances, low-dose, robustness

1 | INTRODUCTION

Deep learning-based methods have revolutionized the field of medical image formation in general and computed tomography (CT) in particular by delivering cutting-edge solutions to a wide range of problems. These include noise reduction,[1–5] image reconstruction,[6–8] scatter estimation,[9–11] and artifact reduction.[12,13] Most of these problems, however, are not injective, meaning that a single target-domain (e.g., artifact-free) image can be derived from different source-domain (e.g., artifact-deteriorated) images.

Therefore, a good network for these tasks *must* be invariant to some input features (e.g., image noise for low-dose reconstruction) to some extent.[14] From a network architecture perspective, invariances can be realized by certain noninjective layers such as max-pooling layers or convolutions with certain weight configurations.

In this study, we aim to investigate and interpret these invariances in low-dose computed tomography (LDCT) image denoising networks — a prevalent application of deep learning in CT image formation. Such an analysis can provide valuable insights into the networks behavior and help in identifying potential biases or shortcomings of the networks and their training data. This is important in order to improve the interpretability and robustness of deep learning-based methods for medical imaging, which is an essential step toward bridging the implementation gap of deep learning-based methods in medical imaging.[15,16]

1.1 | Deep learning-based low-dose CT image denoising

While our method for reconstructing and analyzing invariances of image-to-image translation networks is applicable to a wide range of deep learning-based applications for CT and other modalities, we here focus on the task of LDCT due to the abundance of publications in the field* and the availability of open-source datasets.

LDCT aims at providing an image x with a lower dose than conventional CT acquisitions, which is typically accomplished by decreasing the tube current and consequently reducing the x-ray flux. However, this approach increases noise in the projection data due to photon starvation. As a result, when these images are reconstructed using standard filtered back projection (FBP), they exhibit unwanted noise and streak artifacts, potentially reducing diagnostic value.

To mitigate these artifacts, advanced reconstruction techniques such as iterative reconstruction can be employed. These methods effectively suppress the artifacts but are computationally expensive, often limiting their clinical applicability in time-critical scenarios, such as emergency rooms. On the other hand, denoising methods present a computationally efficient solution and can be integrated seamlessly into any existing reconstruction pipeline. These algorithms may be conventional,[17–20] or data-driven[1,2,21–23] and can be applied in either projection domain, image domain, or both. Particularly, deep learning-based methods applied to reconstructed images are prevalent in the literature since they do not require access to the (often proprietary) projection data.

* For example, PubMed (http://pubmed.ncbi.nlm.nih.gov) lists 56 publications in 2023 for the query: `(low dose OR low-dose) AND (Computed Tomography OR CT) AND deep learning AND denoising`.

Deep learning-based image domain denoising methods usually learn a mapping $f_\theta : x^{(i)} \to y^{(i)}$ from low-dose images $x^{(i)}$ (i.e., images reconstructed from low-dose projections via FBP) to high-dose images $y^{(i)}$, where f_θ is a deep neural network (DNN) with parameters θ. Most methods optimize the parameters in a supervised fashion by minimizing some (typically pixel-wise) loss $\mathcal{L}$ over the training set $\{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$

$$\theta^* = \arg\min_\theta \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(f_\theta(x^{(i)}), y^{(i)}\right). \tag{1}$$

Numerous other works train f_θ unsupervised or self-supervised. These include methods leveraging the image prior of convolutional neural networks (CNNs),[8] intrinsic similarities within the training data (e.g., across views or patches),[5,24–27] or methods from deep metric learning (DML).[28] We refer the reader to Lei et al., 2024[29] for a comprehensive review of these methods.

For a fair comparison between denoising algorithms, we henceforth focus on methods trained using Equation 1 that vary in their architectural design of f_θ and the choice of $\mathcal{L}$ used for learning the parameters θ.

1.2 | Reconstructing invariances of DNNs

Previously, Rombach et al.[14] presented a method to reconstruct the invariances of some image classification network $f : \mathbb{R}^{n\times m} \to \{0, 1\}^c$, with $n \times m$ being the image size and c the number of classes, using conditional invertible neural networks (cINNs). Let $\bar{z} \in \mathbb{R}^d$ denote any internal latent representation (e.g., if $d = n \times m \times 64$ this could be the output of a zero-padded convolutional layer with 64 filters) that we can get by decomposing f into $f(x) = \Psi(\bar{z}) = \Psi(\Phi(x))$, where $\Phi : \mathbb{R}^{n\times m} \to \mathbb{R}^d$ and $\Psi : \mathbb{R}^d \to \{0, 1\}^c$. To then find out which information about x is captured in $\bar{z}$ and which is missing (i.e., the invariances of Φ), we need a compact data representation of x. The authors propose to learn such a data representation z by training a variational autoencoder (VAE) comprised of an encoder E and decoder D. Since $z = E(x)$ now not only contains the information of x that is captured in $\bar{z}$, but also ϕ's invariances v, we need to disentangle these two components. This is achieved by training a normalizing flow $t(\cdot|\bar{z}) : z \to v$ that maps between those two domains, conditioned on the network representation $\bar{z}$. Since t is invertible, we can then sample from $p(v)$ (here assumed to be normal) and apply t^{-1} to obtain samples from $p(z)$. Finally, we can reconstruct the invariances of Φ in image space by applying the pretrained decoder D to the samples z.

This method has later been adapted to reconstruct the invariances of CT image denoising networks.[30] However, due to the fact that LDCT denoising networks

exhibit fewer and more subtle invariances than image classification networks, some reconstructed invariances may be attributed to the VAE rather than the denoising networks. This is further exacerbated by the diversity of medical image data, which makes it difficult for the VAE to learn an almost complete data representation of the input data. In this work, we propose to reconstruct the invariances of LDCT denoising networks by training a *conditional* VAE, therefore improving its data representation compared to previous works. We also investigate the invariances of more recent and advanced denoising networks and introduce methods to analyze the sampled invariances.[†]

2 | METHODS

In the following we will present a method to sample and analyze the invariances of LDCT denoising networks. Note that the method presented herein is network and application-agnostic and therefore potentially applicable to many other image-to-image translation tasks in medical imaging such as metal artifact correction or sparse view CT.

2.1 | Dataset

For all our experiments, we use the 50 chest exams provided in the open-source *Low-dose CT Image and Projection Dataset*.[31] For each scan in the dataset, the authors simulated low-dose reconstructions by inserting noise in the projection domain. These reconstructions correspond to a dose level of 10%.

We randomly split the respective acquisitions (on a patient level) into 70% training (35 patients), 20%validation (10 patients), and 10% (5 patients) test data. During training and validation, we employ a weighted sampling scheme, ensuring that every acquisition has an equal probability of being selected, regardless of the varying number of slices per acquisition. All data are normalized to have zero-mean, unit-variance before feeding them to the networks.

2.2 | Denoising methods

We reconstruct the invariances of four different deep learning-based image denoising algorithms, which are summarized in the following. We refer the reader to the respective publications for more details.

CNN-10[1] One of the earliest deep learning-based methods for LDCT image denoising. The authors propose a simple three-layer CNN which receives low-dose images as input and is trained using Equation 1 with the mean-squared-error loss.

RED-CNN[2] This method builds upon CNN-10 by incorporating a deeper residual encoder-decoder architecture but keeps the overall training procedure identical. In previous works,[32] it has been shown that this method outperforms many other (and notably newer) deep learning-based denoising methods.

WGAN-VGG[3] The authors improve on CNN-10 by using a deeper network architecture and by training it together with a convolutional critic as Wasserstein-GAN.[33] Furthermore, they added a perceptual loss[34] derived from a pretrained VGG to the overall generator loss. In comparison with traditional pixel-wise loss functions, this approach leads to denoised samples that exhibit more refined details and authentic noise textures.

DU-GAN[22] Similar to WGAN-VGG, the authors employ an adversarial training scheme, but use a U-Net-based discriminator[35] which allows for per-pixel feedback to the generator network, for which they use the same structure as RED-CNN.

All four methods are trained using the data as described in Section 2.1 and we use the best performing network on the validation data for subsequent invariance reconstruction. Additional training specifics for each method are provided in Supplementary Materials A.1.

2.3 | Reconstructing invariances

Our pipeline to reconstruct invariances (Figure 1) comprises three components:

(a) The LDCT denoising network f_θ, that receives low-dose images $x^{(i)}$ as input and predicts high-dose images $\hat{y}^{(i)} = f_\theta\left(x^{(i)}\right)$ (Section 2.2).
(b) A conditional VAE $D_\psi \circ E_\varphi$ that is trained to learn a complete data representation $z \in \mathbb{R}^M$ of the low-dose images. We condition both encoder E_φ and decoder D_ψ on predictions of the denoising network $\hat{y}^{(i)}$, thereby improving their encoding/reconstruction capabilities (Section 2.3.1).
(c) A cINN that disentangles the information in z that the denoising network f_θ is invariant to from the one it is not invariant to. To reconstruct invariances, we then sample from the Gaussian distribution of invariances, apply the inverse cINN, and decode the samples using the (fixed) conditional decoder.

2.3.1 | Training of the conditional VAE

In order to reconstruct which information of low-dose images a given denoising network f_θ has learned to represent and which to ignore (i.e., its invariances), we first need to learn an (almost) complete representation of low-dose images $x^{(i)}$. We do so by training a conditional variational autoencoder comprised of a conditional probabilistic encoder E_φ defining the distribution $q_\varphi(z|x, \hat{y})$ and conditional probabilistic decoder

[†] Code available at https://github.com/eeulig/ldct-invariances.

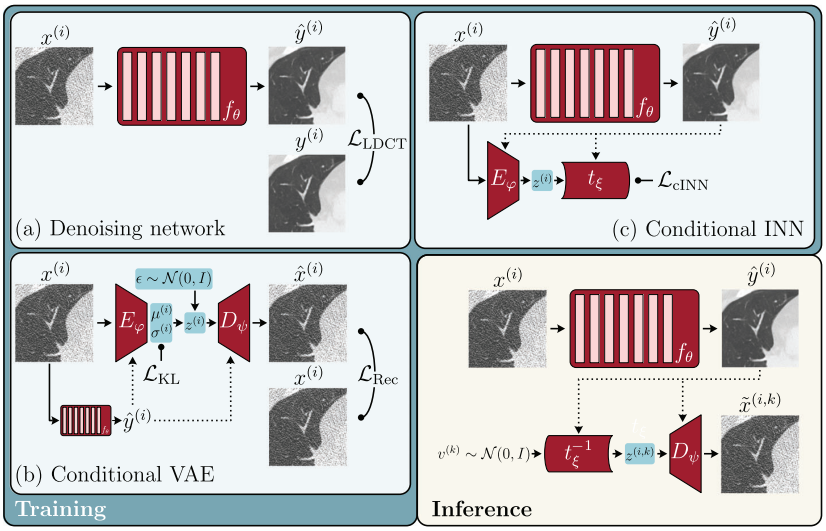


FIGURE 1 Overview of our method to reconstruct invariances of LDCT image denoising networks. Solid arrows represent inputs/outputs to modules, dotted arrows represent conditional inputs to a module. We indicate points where loss functions $\mathcal{L}$ are calculated using $\bullet\!\!-\,\mathcal{L}$. Training: (a) training of the denoising network f_θ using low-dose images $x^{(i)}$ and corresponding high-dose images $y^{(i)}$. $\mathcal{L}_{\text{LDCT}}$ can be some pixel-wise or adversarial loss, or a combination of both (Equation 1); (b) training of the conditional VAE with encoder E_φ and decoder D_ψ conditioned on the denoised images $\hat{y}^{(i)} = f_\theta(x^{(i)})$. $\mathcal{L}_{\text{KL}}$ and $\mathcal{L}_{\text{Rec}}$ are the Kullback–Leibler divergence and reconstruction loss, respectively (Equation 4); and (c) training of the conditional INN t_ξ to disentangle the invariances of the denoising network f_θ from the latent representation $z^{(i)}$ learned by the VAE. $\mathcal{L}_{\text{cINN}}$ is the loss function of the cINN (Equation 6). Inference: We sample new invariances $v^{(k)}$ from the Gaussian distribution of invariances and apply the inverse cINN to obtain samples $z^{(i,k)}$. We then decode the samples using the conditional decoder D_ψ to obtain the invariance reconstructions $\tilde{x}^{(i,k)}$ (Section 2.3).

D_ψ defining $p_\psi(x|z,\hat{y})$. We assume a Gaussian prior $p(z|\hat{y})$ on latent variables z and approximate the posterior with a Gaussian $q_\varphi(z|x,\hat{y})$ with diagonal covariance. Let $\mu^{(i)}, \sigma^{(i)} \in \mathbb{R}^M$ denote the mean and standard deviation predicted by the encoder E_φ for the i-th sample $x^{(i)}$, conditioned on its respective denoised image $\hat{y}^{(i)}$. Then,

$$\ln q_\varphi(z|x^{(i)},\hat{y}^{(i)}) = \ln \mathcal{N}(z;\mu^{(i)},\operatorname{diag}(\sigma^{(i)}))\,. \tag{2}$$

As for any variational autoencoder,[36] both encoder and decoder are trained to maximize the expectation $\mathbb{E}_i\left[\text{ELBO}(x^{(i)},\hat{y}^{(i)})\right]$ with the evidence lower bound (ELBO) being

$$\begin{aligned}\text{ELBO}(x^{(i)},\hat{y}^{(i)}) = \mathbb{E}_{z\sim q_\varphi(z|x^{(i)},\hat{y}^{(i)})}\left[\ln p_\psi(x^{(i)}|z,\hat{y}^{(i)})\right]\\ -D_{\text{KL}}\left(q_\varphi(z|x^{(i)},\hat{y}^{(i)})||p_\psi(z|\hat{y}^{(i)})\right),\end{aligned} \tag{3}$$

where $D_{\text{KL}}(q||p)$ denotes the Kullback–Leibler (KL) divergence between distributions q, p. Using the fact that the KL divergence between two Gaussians can be computed analytically, we derive the loss function

$$\begin{aligned}\mathcal{L}_{\text{VAE}}(\varphi,\psi) &= -\mathbb{E}_i\left[\ln p_\psi(x^{(i)}|z^{(i)},\hat{y}^{(i)})\right.\\ &\quad\left.+\frac{1}{2}\sum_{m=1}^{M}\left(1+\ln(\sigma_m^{(i)})^2-(\mu_m^{(i)})^2-(\sigma_m^{(i)})^2\right)\right]\\ &= -\mathbb{E}_i\left[\underbrace{\|x^{(i)}-\hat{x}^{(i)}\|}_{\mathcal{L}_{\text{Rec}}}+\underbrace{\frac{1}{2}\sum_{m=1}^{M}\left(1+\ln(\sigma_m^{(i)})^2-(\mu_m^{(i)})^2-(\sigma_m^{(i)})^2\right)}_{\mathcal{L}_{\text{KL}}}\right],\end{aligned} \tag{4}$$

where $M = \dim(z)$, $\hat{x}^{(i)} = D_\psi(\mu^{(i)} + \sigma^{(i)}\epsilon|\hat{y}^{(i)})$ with $\epsilon \sim \mathcal{N}(0, I)$.

Conditioning the VAE on auxiliary information[37] eases the task for both encoder and decoder, as they can focus on the information about the input image that is not contained in the auxiliary information (here: the denoised image $\hat{y}$) already[‡].

In our experiments, both E_φ and D_ψ are parameterized by DNNs, with E_φ being an ImageNet-pretrained ResNet-50[38] and D_ψ based on BigGAN.[39] To improve reconstruction quality, we use a perceptual loss[34] and adversarial loss in addition to the pixel-wise loss in Equation 4. We refer the reader to Supplementary Material A.2 for more details on the training procedure. For comparison, we also train a VAE without the conditioning on $\hat{y}$ (as explored in previous works[14,30]),

[‡] Note, that this has the interesting side-effect that the latent space is already mostly comprised of the invariances of the denoising network as this is exactly the information missing in $\hat{y}$.

but otherwise identical architecture and training procedure.

2.3.2 | Training of the conditional invertible neural network

The latent representation z does not only contain invariances of the denoised image $\hat{y}^{(i)}$ but also information about the input image x. Therefore, we need to disentangle these two components, that is, extract the invariances v of the denoising networks' prediction $\hat{y}^{(i)}$ from the other information in z. Thus, we need to learn a mapping t_ξ from z to some space of invariances, *given* a denoised image $\hat{y}^{(i)}$. Let this space of invariances $p(v)$ be a standard Gaussian distribution, that is, $p(v) \sim \mathcal{N}(0, I)$. Then, $t_\xi : p(z|\hat{y}) \to p(v)$ allows us to generate $v^{(i)} = t_\xi(z^{(i)}|\hat{y}^{(i)})$ for any given sample i. In our experiments, t_ξ is realized by a conditional invertible neural network with parameters ξ, that is, a normalizing flow conditioned on $\hat{y}^{(i)}$.[40–43]

As for any cINN,[43] we can find optimal parameters ξ via standard maximum likelihood training. Using the change-of-variables formula gives us the likelihood

$$q(z^{(i)}|\hat{y}^{(i)}, \xi) = p\big(t_\xi(z^{(i)}|\hat{y}^{(i)})\big)\left|J^{(i)}\right|, \tag{5}$$

with $J^{(i)} = \det\left(\frac{\partial t_\xi(z^{(i)}|\hat{y}^{(i)})}{\partial z^{(i)}}\right)$. The loss function over training samples i then reads as

$$\begin{aligned}\mathcal{L}_{\text{cINN}} &= \mathbb{E}_i\left[-\ln q(z^{(i)}|\hat{y}^{(i)}, \xi)\right] \\ &= \mathbb{E}_i\left[\underbrace{-\ln p\big(t_\xi(z^{(i)}|\hat{y}^{(i)})\big)}_{\ell(t_\xi(z^{(i)}|\hat{y}^{(i)});0,1)} - \ln\left|J^{(i)}\right|\right] \\ &= \mathbb{E}_i\left[\frac{1}{2}\|t_\xi(z^{(i)}|\hat{y}^{(i)})\|_2^2 - \ln\left|J^{(i)}\right|\right],\end{aligned} \tag{6}$$

where in the last step we used the log-likelihood over samples $x = \{x_1, x_2, \dots, x_N\}$ of a standard Gaussian distribution $\ell(x; \mu, \sigma^2) = -N \ln\sigma - \frac{N}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\|x-\mu\|_2^2$ and the assumption that $p(v)$ is a normal distribution with zero mean and unit variance. The first line in Equation 6 is the log-likelihood of observing some representation $z^{(i)}$ given the corresponding denoised image $\hat{y}^{(i)}$ under parameters ξ. After optimization of parameters ξ, we can sample from $p(v)$ and apply the inverse t_ξ^{-1} to map invariances to the input data representation, conditioned on the denoised image $\hat{y}^{(i)}$. We refer the reader to Supplementary Material A.3 for more details on the architecture and training of t_ξ.

2.3.3 | Sampling invariances

Once conditional VAE and cINN are trained, we can generate invariance samples $\tilde{x}^{(i,k)}$ for a given sample $x^{(i)}$ from the test set and a (trained) denoising network f_θ as follows:

1. Denoise the image using the pretrained denoising network f_θ: $\hat{y}^{(i)} = f_\theta(x^{(i)})$.
2. Sample $v^{(k)} \sim \mathcal{N}(0, I)$ from the space of invariances.
3. Apply the inverse of the cINN t_ξ^{-1} to the sampled invariance: $z^{(i,k)} = t_\xi^{-1}(v^{(k)}|\hat{y}^{(i)})$.
4. Decode the samples using the (fixed) conditional decoder D_ψ to obtain the invariance reconstructions $\tilde{x}^{(i,k)} = D_\psi(z^{(i,k)}|\hat{y}^{(i)})$.

Every $\tilde{x}^{(i,k)}$ is then a sample from the distribution of invariances of the denoising network f_θ for the i^{th} low-dose image $x^{(i)}$ and two images $\tilde{x}^{(i,k)}, \tilde{x}^{(i,l)}$ differ only in their realization of invariances.

2.4 | Analyzing invariances

In our experiments, we find that the most prominent invariances of LDCT denoising networks are related to the noise level and noise realization of the input images. While this is expected and a desirable property of any denoising algorithm, this does not answer our initial question of whether LDCT denoising networks are invariant to anatomical structures or other image *content*. Finding such differences in the pixel space is challenging, as differences in noise realizations can easily overshadow differences in content. We therefore propose to analyze the invariances in an embedding space instead and compare two different methods to do so (Figure 2). The first is based on an embedding learned by an unconditional VAE (Sections 2.4.1, and Figure 2a). The second is based on a learned embedding of the invariances using a DML approach (Sections 2.4.2, and Figure 2b).

2.4.1 | Using an unconditional VAE

The conditional VAE is trained to learn a complete representation of the low-dose images, which includes both their noise level and noise realization as well as the anatomical structures and other image content. However, since for the conditional VAE, the latent space follows the distribution $p(z|\hat{y}^{(i)})$, we cannot compare different samples i with another. Instead, we use an unconditional VAE for which $p(z)$ is standard Gaussian distributed but which is otherwise identical to the conditional one. Since noise is generally harder to model than content, we expect the learned

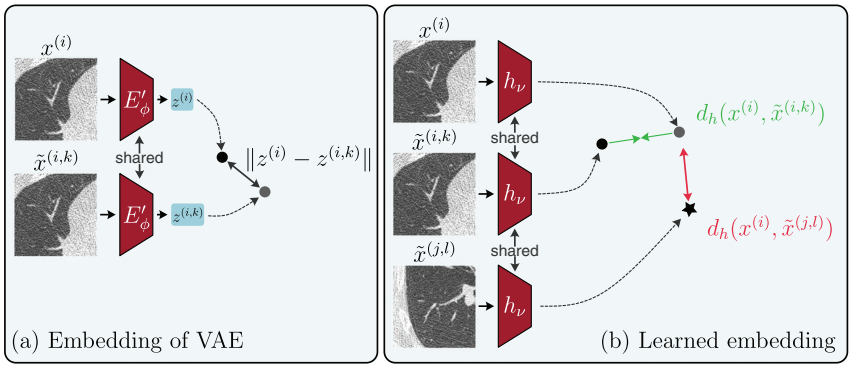


FIGURE 2 Overview of our methods to analyze sampled invariances. (a) based on the embedding of an unconditional VAE, with encoder E'_ϕ, whose latent space is dominated by content-related information. Applying the same encoder to the input image $x^{(i)}$ and sampled invariance $\tilde{x}^{(i,k)}$, we can measure the content similarity between the two and (b) based on a learned embedding which is trained with a triplet loss (Equation 8) to map low-dose images $x^{(i)}$ closer to invariance samples $\tilde{x}^{(i,k)}$ corresponding to the same sample i than to invariance samples $\tilde{x}^{(j,l)}, j \neq i$.

representation $E'_\phi(x^{(i)})$ to be dominated by content-related information.

We can then use differences in the latent space as a proxy for differences in anatomical content between invariance samples $\tilde{x}^{(i,k)}$ and low-dose inputs $x^{(i)}$. To this end we compute the cosine similarity between the latent representations of the low-dose input and invariance samples $z^{(i)} = E'_\phi(x^{(i)})$, $z^{(i,k)} = E'_\phi(\tilde{x}^{(i,k)})$ as

$$\cos\left(z^{(i)}, z^{(i,k)}\right) = \frac{z^{(i)} \cdot z^{(i,k)}}{\|z^{(i)}\| \|z^{(i,k)}\|}. \tag{7}$$

2.4.2 | Using a learned embedding

We can also learn an embedding of the invariances using a DML approach. Metric learning generally seeks to learn a metric function h such that semantic relations between datapoints $x^{(i)}, x^{(j)} \in \mathcal{X}$ are depicted by metric distances $d_h(x^{(i)}, x^{(j)}) := d(h(x^{(i)}), h(x^{(j)}))$, with $d(\cdot, \cdot)$ being some distance, in the embedding $h(\cdot)$. In DML, h is typically parameterized by a deep neural network $h_\nu := h$ with weights ν being learned by minimizing a loss function that encourages the network to map similar (w.r.t. some semantic relation) samples closer together than dissimilar ones. To do so, many different loss functions have been proposed, most popularly ranking-based loss functions.[44–46] We refer the reader to Roth et al., 2020[47] for a nice overview of training strategies in DML.

In our experiments, we use the triplet loss[45] to learn an embedding in which low-dose inputs $x^{(i)}$ are closer to invariance samples $\tilde{x}^{(i,k)}, \forall k$ corresponding to the same sample i (with same anatomy) than to invariances samples $x^{(j,l)}, \forall l, j \neq i$ corresponding to different samples (with different anatomy). The loss function for h_ν then reads as

$$\mathcal{L}_{\text{Triplet}} = \mathbb{E}_{i,j}\left[d_h\left(x^{(i)}, \tilde{x}^{(i,k)}\right) - d_h\left(x^{(i)}, \tilde{x}^{(j,l)}\right) + \alpha\right]_+, \tag{8}$$

with α being some prespecified margin. In our experiments, we use a pretrained ResNet-50[38] as h_ν and select triplets $(a, p, n) := (x^{(i)}, \tilde{x}^{(i,k)}, \tilde{x}^{(j,l)})$ using the semi-hard triplet mining strategy.[45] We refer the reader to Supplementary Material A.4 for more details on the training procedure.

3 | RESULTS

3.1 | Denoising of LDCT images

We first verify qualitatively that the denoising methods (compare Section 2.2) perform as expected and are able to denoise LDCT images similarly as reported in their respective publications. To this end, we show results for random axial slices of all five test patients in Figure 3. For each patient of the test set, we show the high-dose image, the low-dose image, and the respective denoised images. While all methods are able to reduce noise and streak artifacts compared to the low-dose image, the results of WGAN-VGG[3] and DU-GAN[22] show more realistic noise structures and exhibit finer details compared to the two methods trained using a pixel-wise loss exclusively. This is in line with the findings of Yang et al.[3] and Huang et al.[22] and can be attributed to the additional perceptual loss (for WGAN-VGG) and adversarial loss (for both WGAN-VGG and DU-GAN).

Upon quantitative evaluation (Table 1), we find that RED-CNN performs best in terms of the structural similarity index measure (SSIM), peak signal-to-noise ratio (PSNR), and root-mean-square error (RMSE). However, it is important to note that these metrics do not correlate well with human reader ratings (the gold

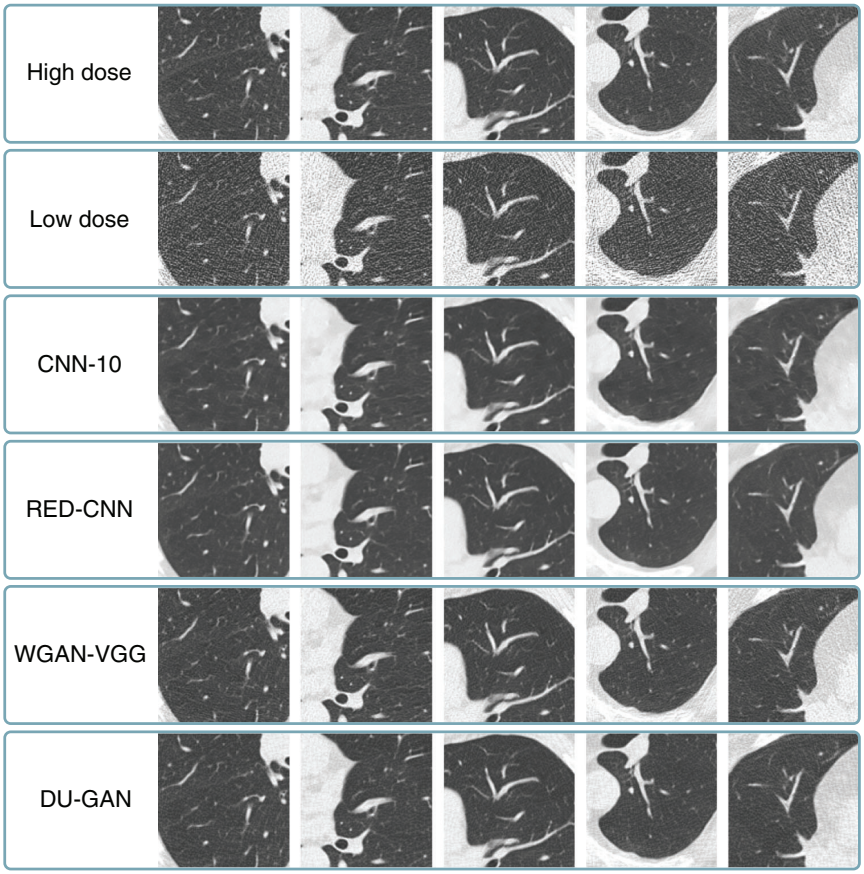


FIGURE 3 High-dose, low-dose, and denoising results for the four methods described in Section 2.2. We show results for random axial slices and crops of size 128 × 128 px for all five patients from the test set. Center (C) and width (W) are $C = -600$ HU, $W = 1500$ HU.

TABLE 1 Quantitative evaluation of the denoising methods described in Section 2.2.

	SSIM	PSNR (dB)	RMSE (HU)
LD	0.312 ± 0.072	18.1 ± 2.5	236 ± 86
CNN-10	0.56 ± 0.10	27.3 ± 2.1	72 ± 19
RED-CNN	**0.58 ± 0.10**	**28.0 ± 2.2**	**66 ± 18**
WGAN-VGG	0.505 ± 0.099	25.3 ± 2.2	91 ± 26
DU-GAN	0.544 ± 0.096	26.3 ± 2.2	80 ± 22

Note: We report the mean and standard deviation of the SSIM, PSNR, and RMSE over all axial slices of the test set. **Bold** values highlight the best performing method for each metric.

standard in terms of medical image quality assessment) for computed tomography.[48–50] Since this work is not concerned with the evaluation of the denoising methods themselves, but rather with their invariances, we do not further investigate the performance of the denoising methods and leave the development of better metrics for future work.

3.2 | VAE reconstructions

Next, we evaluate the reconstruction capabilities of the conditional VAE (Section 2.3.1) for random axial slices of all patients of the test set in Figure 4. We find that reconstructions of the conditional VAE (Figure 4; third row) are very similar to the input low-dose images (Figure 4; second row) for all exam types. Additionally, we show the reconstructions of an unconditional VAE (Figure 4; last row) as it was used in previous work to reconstruct invariances of LDCT denoising networks[30] for comparison. While the unconditional VAE is able to generate realistic low-dose images that reflect, to some extent, the anatomical structures of the low-dose input images, it fails to capture fine details and removes or hallucinates many of the anatomical structures in the reconstructions (compare red arrows in Figure 4). We show reconstruction results for all conditional VAEs (conditioned on different denoising networks) in Supplementary Material B.

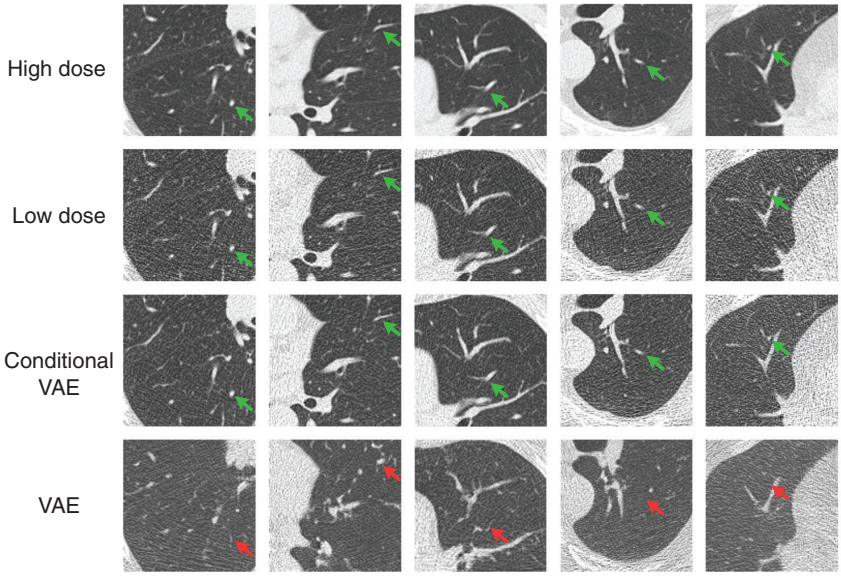


FIGURE 4 High-dose, low-dose, and VAE reconstructions for the conditional VAE (here conditioned on RED-CNN) described in Section 2.3.1. Patients, axial slices and crops correspond to those shown in Figure 3. Additionally, we show the reconstructions of an unconditional VAE (as used in previous work[30]) for comparison. $C = -600$ HU, $W = 1500$ HU.

3.3 | Invariance reconstruction

Given the procedure described in Section 2.3.3 and shown in Figure 1, we sample 100 invariances for each of the four denoising networks on 1000 random crops of the test set. In Figure 5 we show three invariances for one of those random crops. We find that for all denoising networks sampled invariances mainly differ in terms of noise amplitude and realization. This is expected as the networks see many noise realizations as well as patients of different thickness (influencing the noise level) during the training. These differences in noise structure and amplitude overshadow possible differences in anatomical content between samples. We provide additional results in Supplementary Material B.

3.4 | Analyzing invariances

In the VAE latent space

Next, we analyze sampled invariances in the latent space of the unconditional VAE. To this end we compute for each network and sampled crop i the mean cosine similarity over sampled invariances $k = 1, 2, \dots, K$

$$S^{(i)}_{\text{VAE}} = \frac{1}{K}\sum_{k=1}^{K} \cos\left(z^{(i)}, z^{(i,k)}\right). \tag{9}$$

In Figure 6 we show four crops, corresponding to the $\{0, 1/3, 2/3, 1\}$ quantiles of the mean similarity $S^{(i)}_{\text{VAE}}$ over all test samples, for each network. Here we find that for samples with lower $S^{(i)}_{\text{VAE}}$ (left), most differences between $x^{(i)}$ and $\tilde{x}^{(i,k)}$ are in terms of anatomical content (red arrows). In contrast, for samples with higher $S^{(i)}$ (right) anatomical content is similar between $x^{(i)}$ and $\tilde{x}^{(i,k)}$ and differences are mainly in terms of noise amplitude and realization. This indicates that the latent space of the VAE is indeed dominated by anatomy-related information and disentangling sampled invariances. We provide further analysis of the VAE latent space in Supplementary Material B.

In the DML latent space

Next, we analyze sampled invariances using the DML-based embedding. Similar as for the VAE, we measure similarity between $x^{(i)}$ and $\tilde{x}^{(i,k)}$ using the mean cosine similarity over sampled invariances $k = 1, 2, \dots, K$

$$S^{(i)}_{\text{DML}} = \frac{1}{K}\sum_{k=1}^{K} \cos\left(h(x^{(i)}), h(\tilde{x}^{(i,k)})\right) \tag{10}$$

and show four samples with increasing mean similarity $S^{(i)}_{\text{DML}}$, again corresponding to the $\{0, 1/3, 2/3, 1\}$ quantiles of the empirical distribution, in Figure 7. For all denoising networks, we find that samples with lower $S^{(i)}_{\text{DML}}$ (left) exhibit differences in terms of anatomical content (red arrows) while samples with higher $S^{(i)}_{\text{DML}}$ (right) mainly differ in terms of noise amplitude and realization.

Lastly, we compare the invariances of different denoising networks quantitatively using the pixel-wise

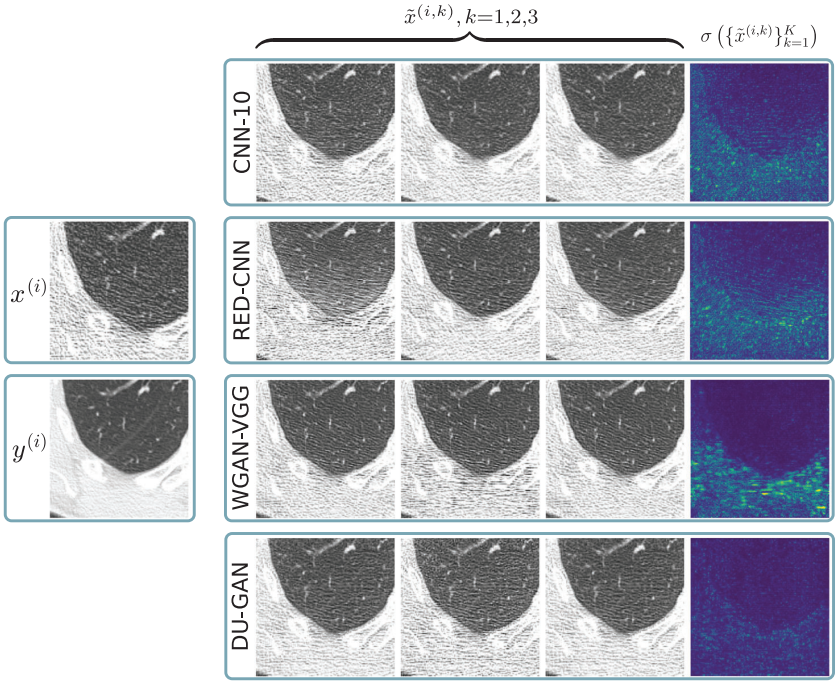


FIGURE 5 Invariances for a random crop i from the test set. Shown are low-dose image $x^{(i)}$, high dose image $y^{(i)}$ and three reconstructed invariances $\tilde{x}^{(i,k)}$ for each of the four denoising methods. We also show standard deviations $\sigma\left(\{\tilde{x}^{(i,k)}\}_{i=1}^{K}\right)$ over $K = 100$ invariances. For CT Images: $C = -600$ HU, $W = 1500$ HU, for standard deviations: $C = 0$ HU, $W = 400$ HU.

TABLE 2 Quantitative evaluation of invariances using the mean absolute difference (MD), mean cosine similarity in the VAE latent space (S_{VAE}), and mean cosine similarity in the learned embedding space (S_{DML}).

Invariances	MD ↑ noise + content	S_{VAE} ↓ content	S_{DML} ↓ content
CNN-10	$182 \pm 67^{***}$	$0.978 \pm 0.020^{**}$	$0.997 \pm 0.004^{***}$
RED-CNN	$\mathbf{191 \pm 67}$	$\mathbf{0.976 \pm 0.022}$	$\mathbf{0.996 \pm 0.007}$
WGAN-VGG	$158 \pm 64^{***}$	$0.979 \pm 0.021^{***}$	$\mathbf{0.996 \pm 0.006}$
DU-GAN	$178 \pm 70^{***}$	0.979 ± 0.013	0.997 ± 0.004

Note: For MD, higher values imply more noise and content-related invariances (due to MD being pixel-wise), for S_{VAE} and S_{DML} lower (↓) values imply more anatomical invariances (since they measure similarity of anatomical content between sampled invariances and input images). **Bold** values indicate the denoising method with the highest amount of invariances (↑ MD, ↓ S_{VAE}, ↓ S_{DML}). We indicate statistical significance of this finding with $^{*}(p < 0.05)$, $^{**}(p < 0.01)$, and $^{***}(p < 0.001)$.

mean absolute difference (MD) between $x^{(i)}$ and $\bar{x}^{(i,k)}$ as well as $S_{\text{VAE}}^{(i)}$ and $S_{\text{DML}}^{(i)}$ (Table 2). Note that, opposed to S_{VAE} and S_{DML}, the MD acts in the pixel space and is therefore both a measure of content-related and noise-related invariances. We find that quantitatively, the invariances of the denoising networks are very similar, with RED-CNN showing the highest amount of invariances (higher MD, lower S_{VAE} and S_{DML}). Upon a statistical analysis using a one-sided Mann–Whitney U test with Benjamini–Hochberg correction for multiple comparisons, we examine that this finding is significant for most invariance metrics and denoising methods (In Table 2, stars for some method indicate significance levels of the pairwise test that RED-CNN has more invariances compared to this method).

4 | DISCUSSION

In this work, we presented a method for reconstructing the invariances of deep learning-based low-dose CT image denoising algorithms. Upon reconstructing the invariances of four common denoising networks we found that the sampled invariances mainly differ in terms of noise amplitude and realization, while the anatomical content is largely preserved. This is expected and can be explained by the training procedure of these networks. To answer our initial question of whether LDCT denoising networks are invariant to anatomical structures or other image content, we further proposed two methods to analyze the sampled invariances. Both methods are based on measuring distances between sampled invariances and input images in a lower-dimensional latent space. Using these methods, we found that all denoising networks are also invariant to anatomical structures to some extent. Quantitatively, the amount of

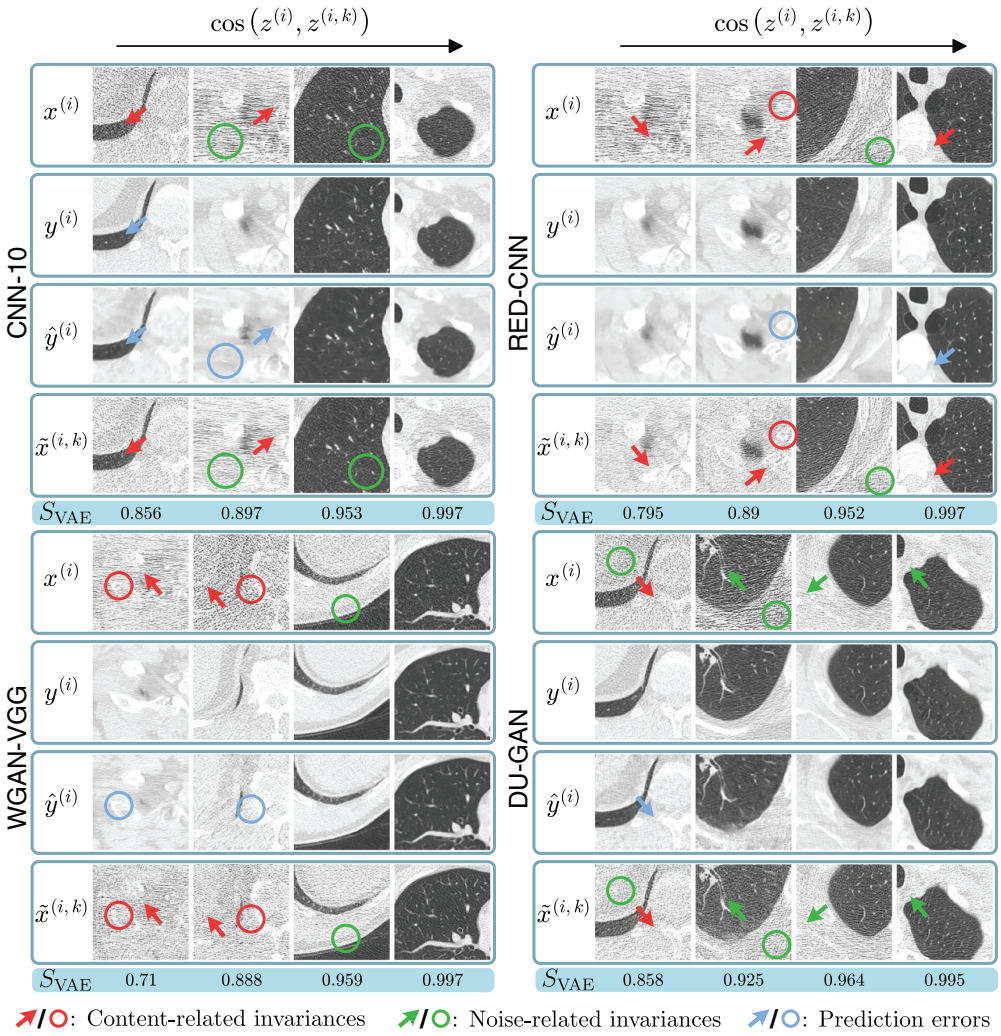


FIGURE 6 Invariances with increasing S_{VAE} (left to right), that is, decreasing amount of content-related invariances as measured in the VAE latent space, for each of the four denoising methods. $C = -600$ HU, $W = 1500$ HU.

invariances (both noise-related and content-related) of different denoising networks are very similar, with RED-CNN showing the highest amount of invariances both in terms of noise and anatomical structures. In Supplementary Material B we provide additional results for an algorithm that has, by design, more invariances to anatomical structures.

Our method is similar to uncertainty quantification methods such as Monte-Carlo dropout or moment propagation[51,52] in that it can improve the interpretability of deep learning-based methods for medical imaging. However, both approaches provide orthogonal views of the network's behavior. While uncertainty quantification methods provide a measure of the network's confidence in its predictions, our method provides a measure of the network's invariances to the input features. There are many scenarios in which an algorithm can be confident in its prediction but still exhibit invariances to certain input features (e.g., the algorithm analyzed in Supplementary Material B; *Case study: Algorithm with strong invariances by design*). In such cases, our method can provide additional insights into the network's behavior. Lastly, our proposed approaches for analyzing the sampled invariances could also be helpful in analyzing systematic uncertainties quantified using the aforementioned methods, an interesting direction for future work.

5 | CONCLUSIONS

Our work shows that common LDCT image denoising networks are invariant to certain input features. While these invariances are mostly dominated by noise, all

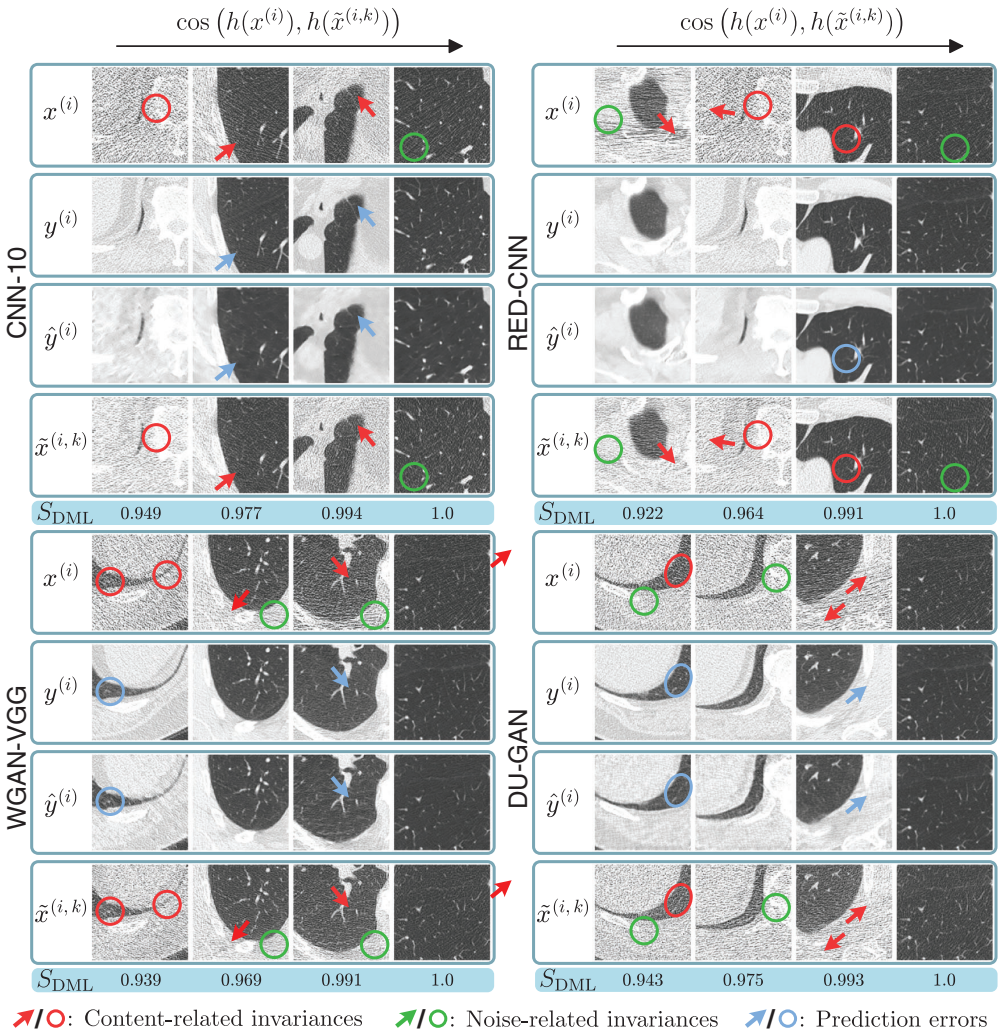


FIGURE 7 Invariances with increasing S_{DML} (left to right), that is, decreasing amount of content-related invariances as measured in the DML embedding space, for each of the four denoising methods. $C = -600$ HU, $W = 1500$ HU.

networks investigated in this study are also invariant to anatomical structures to some extent. We believe that developing methods to reconstruct and analyze these invariances is an important step toward interpreting deep learning-based methods for medical image formation.

Since the presented method is architecture agnostic, several natural extensions of our work come to mind: Promising research directions include (a) evaluating the impact of training data distribution on the invariances of LDCT denoising networks; (b) investigating invariances of other networks for medical imaging including other modalities such as PET and MR; and (c) relating invariances to the similar concept of hallucinations in medical imaging. Lastly, while the sampling of invariances using our method is very fast (≈ 23 ms), future work should reduce the computational complexity of training the two networks, for example by disentangling the invariances in the VAE latent space directly, thus eliminating the need for training of a cINN.

ACKNOWLEDGMENTS

This work was supported in part by the Helmholtz International Graduate School for Cancer Research, Heidelberg, Germany.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

REFERENCES

1. Chen H, Zhang Y, Zhang W, et al. Low-dose CT via convolutional neural network. *Biomed Opt Express*. 2017;8:679-694.

2. Chen H, Zhang Y, Kalra MK, et al. Low-dose CT with a residual encoder-decoder convolutional neural network. *IEEE Trans Med Imaging*. 2017;36:2524-2535.
3. Yang Q, Yan P, Zhang Y, et al. Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. *IEEE Trans Med Imaging*. 2018;37:1348-1357.
4. Wu D, Gong K, Kim K, Li Q. Consensus neural network for medical imaging denoising with only noisy training samples. Medical Image Computing and Computer Assisted Intervention (MICCAI); 2019:741-749. doi:10.1007/978-3-030-32251-9_81
5. Wang S, Yang Y, Yin Z, Wang AS. Noise2Noise for denoising photon counting CT images: generating training data from existing scans. In: *Medical Imaging 2023: Physics of Medical Imaging*. Vol 12463. SPIE; 2023:15-19.
6. Würfl T, Hoffmann M, Christlein V, et al. Deep learning computed tomography: learning projection-domain weights from image domain in limited angle problems. *IEEE Trans Med Imaging*. 2018;37:1454-1463.
7. Huang Y, Preuhs A, Lauritsch G, Manhart M, Huang X, Maier A. Data consistent artifact reduction for limited angle tomography with deep learning prior. In: *Machine Learning for Medical Image Reconstruction: Second International Workshop, MLMIR 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings*, Berlin, Heidelberg, Springer-Verlag; 2019:101-112.
8. Baguer DO, Leuschner J, Schmidt M. Computed tomography reconstruction using deep image prior and learned reconstruction methods. *Inverse Prob*. 2020;36:094004.
9. Maier J, Eulig E, Vöth T, et al. Real-time scatter estimation for medical CT using the deep scatter estimation: method and robustness analysis with respect to different anatomies, dose levels, tube voltages, and data truncation. *Med Phys*. 2019;46:238-249.
10. Hansen DC, Landry G, Kamp F, et al. ScatterNet: A convolutional neural network for cone-beam CT intensity correction. *Med Phys*. 2018;45:4916-4926.
11. Roser P, Birkhold A, Preuhs A, et al. X-ray scatter estimation using deep splines. *IEEE Trans Med Imaging*. 2021;40:2272-2283.
12. Lin W-A, Liao H, Peng C, et al. DuDoNet: dual domain network for CT metal artifact reduction. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2019:10512-10521.
13. Ghani MU, Karl WC. Fast enhanced CT metal artifact reduction using data domain deep learning. *IEEE Trans Comput Imaging*. 2020;6:181-193.
14. Rombach R, Esser P, Ommer B. Making sense of CNNs: interpreting deep representations & their invariances with INNs. In: *European Conference on Computer Vision (ECCV)*. IEEE; 2020:18.
15. Cabitza F, Campagner A, Balsano C. Bridging the "last mile" gap between AI implementation and operation: "data awareness" that matters. *Ann Transl Med*. 2020;8:501.
16. Chen H, Gomez C, Huang C-M, Unberath M. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. *npj Digital Med*. 2022;5:1-15.
17. Manduca A, Yu L, Trzasko JD, et al. Projection space denoising with bilateral filtering and CT noise modeling for dose reduction in CT. *Med Phys*. 2009;36:4911-4919.
18. Balda M, Hornegger J, Heismann B. Ray contribution masks for structure adaptive sinogram filtering. *IEEE Trans Med Imaging*. 2012;31:1228-1239.
19. Feruglio PF, Vinegoni C, Gros J, Sbarbati A, Weissleder R. Block matching 3D random noise filtering for absorption optical projection tomography. *Phys Med Biol*. 2010;55:5401-5415.
20. Li Z, Yu L, Trzasko JD, et al. Adaptive nonlocal means filtering based on local noise level for CT denoising. *Med Phys*. 2014;41:011908.
21. Heinrich MP, Stille M, Buzug TM. Residual U-net convolutional neural network architecture for low-dose CT denoising. *Curr Dir Biomed Eng*. 2018;4:297-300.
22. Huang Z, Zhang J, Zhang Y, Shan H. DU-GAN: generative adversarial networks with dual-domain U-Net-based discriminators for low-dose CT denoising. *IEEE Trans Instrum Meas*. 2022;71:1-12.
23. Shan H, Padole A, Homayounieh F, et al. Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction. *Nat Mach Intell*. 2019;1:269-276.
24. Yuan N, Zhou J, Qi J. Half2Half: deep neural network based CT image ddenoising without independent reference data. *Phys Med Biol*. 2020;65:215020.
25. Zainulina E, Chernyavskiy A, Dylov DV. No-reference denoising of low-dose CT projections. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE; 2021:77-81.
26. Hong Z, Zeng D, Tao X, Ma J. Learning CT projection denoising from adjacent views. *Med Phys*. 2023;50:1367-1377.
27. Niu C, Li M, Fan F, Wu W, Guo X, Lyu Q, Wang G. Noise suppression with similarity-based self-supervised deep learning. *IEEE Trans Med Imaging*. 2023;42:1590-1602.
28. Jung C, Lee J, You S, Ye JC. Patch-wise deep metric learning for unsupervised low-dose CT denoising. In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S, eds. *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Lecture Notes in Computer Science, Cham, Springer Nature Switzerland; 2022:634-643.
29. Lei Y, Niu C, Zhang J, Wang G, Shan H. CT image denoising and deblurring with deep learning: current status and perspectives. *IEEE Trans Radiat Plasma Med Sci*. 2024;8:153-172.
30. Eulig E, Ommer B, Kachelrieß M. Reconstructing invariances of CT image denoising networks using invertible neural networks. In: *International Conference on Image Formation in X-Ray Computed Tomography*. Vol 12304. SPIE; 2022:169-173.
31. McCollough C, Chen B, Holmes III DR, et al. Low dose CT image and projection data (data set). The Cancer Imaging Archive; 2020. doi:10.7937/9NPB-2637
32. Eulig E, Ommer B, Kachelrieß M. Benchmarking deep learning-based low-dose CT image denoising algorithms. arXiv preprint. 2024. 10.1002/mp.17379
33. Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: *International Conference on Machine Learning (ICML)*. PMLR; 2017:214-223.
34. Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution. In: Leibe B, Matas J, Sebe N, Welling M, eds. *European Conference on Computer Vision (ECCV)*. Lecture Notes in Computer Science, Cham, Springer International Publishing; 2016:694-711.
35. Schonfeld E, Schiele B, Khoreva A. A U-Net based discriminator for generative adversarial networks. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA; 2020:8204-8213.
36. Kingma DP, Welling M. Auto-encoding variational Bayes. In: *International Conference on Learning Representations (ICLR)*. 2014.
37. Sohn K, Lee H, Yan X. Learning structured output representation using deep conditional generative models. In: *Advances in Neural Information Processing Systems*. Vol 28. Curran Associates, Inc.; 2015.
38. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, NV, USA; 2016:770-778.
39. Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis. In: *International Conference on Learning Representations (ICLR)*. 2018.
40. Dinh L, Krueger D, Bengio Y. NICE: Non-linear independent components estimation. In: *International Conference on Learning Representations (ICLR), Workshop Track*. 2015.

41. Dinh L, Sohl-Dickstein J, Bengio S. Density estimation using real NVP. In: *International Conference on Learning Representations (ICLR)*. 2017.
42. Rezende DJ, Mohamed S. Variational inference with normalizing flows. In: *International Conference on Machine Learning (ICML)*. ICML'15, Lille, France; 2015:1530-1538. JMLR.org.
43. Ardizzone L, Kruse J, Lüth C, Bracher N, Rother C, Köthe U. Conditional invertible neural networks for diverse image-to-image translation. In: Akata Z, Geiger A, Sattler T, eds. *Pattern Recognition*, Lecture Notes in Computer Science, Springer International Publishing, Cham; 2021:373-387.
44. Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol 2. IEEE; 2006:1735-1742.
45. Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Boston, MA, USA; 2015:815-823.
46. Sohn K. Improved deep metric learning with multi-class N-pair loss objective. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol 29. Curran Associates, Inc.; 2016.
47. Roth K, Milbich T, Sinha S, Gupta P, Ommer B, Cohen JP. Revisiting training strategies and generalization performance in deep metric learning. In: *International Conference on Machine Learning (ICML)*. PMLR; 2020:8242-8252.
48. Verdun FR, Racine D, Ott JG, et al. Image quality in CT: from physical measurements to model observers. *Physica Med*. 2015;31:823-843.
49. Renieblas GP, Nogués AT, Md AMG, León NG, del Castillo EG. Structural similarity index family for image quality assessment in radiological images. *J Med Imaging*. 2017;4:035501.
50. Ohashi K, Nagatani Y, Yoshigoe M, et al. Applicability evaluation of full-reference image quality assessment methods for computed tomography images. *J Digit Imaging*. 2023;36:2623-2634.
51. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: *Proceedings of The 33rd International Conference on Machine Learning*. PMLR; 2016:1050-1059.
52. Liu SZ, Vagdargi P, Jones CK, et al. One-shot estimation of epistemic uncertainty in deep learning image formation with application to high-quality cone-beam CT reconstruction. In: *Medical Imaging 2024: Physics of Medical Imaging*. Vol. 12925. SPIE; 2024:223-228.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Eulig E, Jäger F, Maier J, Ommer B, Kachelrieß M. Reconstructing and analyzing the invariances of low-dose CT image denoising networks. *Med Phys*. 2024;1-13. https://doi.org/10.1002/mp.17413

Chapter 4

Synthetic Training Data for Deep Learning-Based Digital Subtraction Angiography

L. Duan, E. Eulig, M. Knaup, R. Adamus, M. Lell, and M. Kachelrieß. "Training of a Deep Learning Based Digital Subtraction Angiography Method Using Synthetic Data". In: *Medical Physics* 51.7 (2024), pp. 4793–4810

We will now present our work on synthetic training data for deep learning-based methods to predict DSA-like images. After a brief summary in Sec. 4.1 the full paper [72] is provided in Sec. 4.2.

4.1 Summary

4.1.1 Introduction

DSA is a fluoroscopy method that leverages radiographic subtraction to diagnose various cardiovascular diseases. It is primarily used to diagnose arterial and venous occlusions, stenosis, and aneurysms. This includes acute limb ischemia [73], arteriovenous malformations, intracranial aneurysms [74], and renal artery stenosis [75]. During the exam, a series of X-ray images is acquired using a C-arm system while injecting contrast media into the vessels. The contrast media enhances or reduces radiodensity in the vessels, allowing for a clear (positive or negative) contrast to the surrounding tissue in the X-ray image. To selectively display the vessels, a subtraction image is created by subtracting a mask image (X-ray image acquired prior to the injection of contrast media) from the subsequent images taken during the exam and with contrast media.

However, a limitation of DSA is its reliance on static data. Any motion caused by the C-arm system or patient movement leads to artifacts in the subtraction image, thereby diminishing its clinical value. In particular cardiac, pelvis, or abdomen exams are prone to motion artifacts and

thus not suitable for DSA (see Fig. 4.1 for examples of acquisitions with strong motion artifacts in the DSA images).

To overcome these drawbacks, multiple works investigated the use of deep learning-based methods to predict DSA-like images from standard X-ray images [76, 77, 78, 79]. To this end, the networks are trained to predict the DSA image from an X-ray image on static acquisitions, without or with minimal motion artifacts. Trained networks can then be applied to arbitrary contrast-enhanced acquisitions. To remain in-distribution, however, we must have access to static data from the desired anatomical region. While this is feasible for some applications, e.g., bolus chase studies of peripheral angiography, for other applications, e.g., cardiac or abdominal DSA, such static data are not available and the application of networks trained on different anatomies would pose an out-of-distribution (OOD) problem.

In this work, we addressed the problem of training deep learning-based methods for DSA on anatomies where static data are not available. To this end, we proposed a method to generate synthetic training data for DSA by combining simulated vessel structures with forward projections of CT acquisitions. Using our synthetic data, we trained different networks previously proposed for this task and compared their performance to networks trained on clinical DSA data.

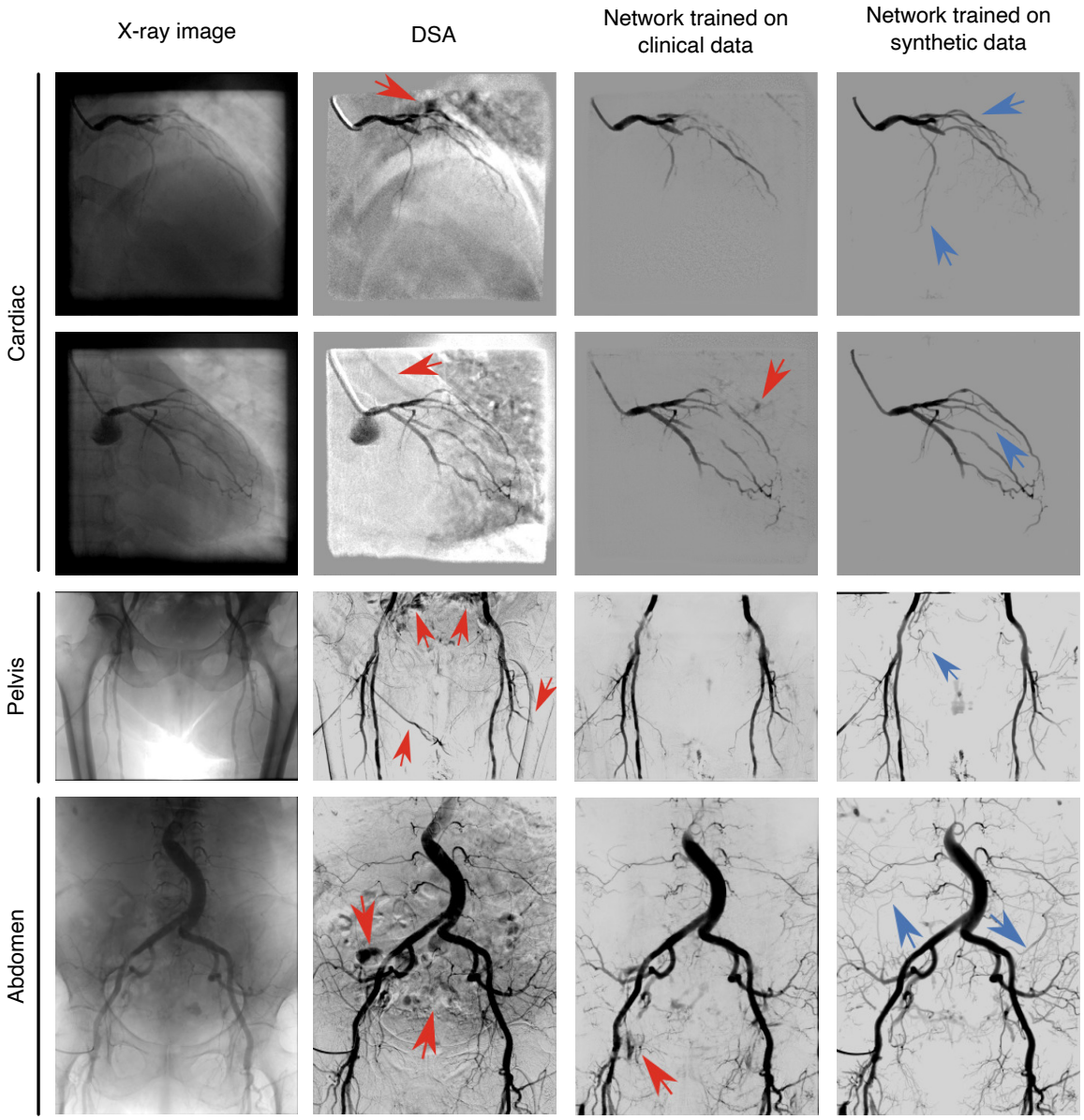
4.1.2 Methods

We simulated vessel skeletons using stochastic Lindenmayer systems (L-systems) [80, 81]. By controlling the parameters of the L-systems, we generated a variety of vessel structures, resembling vessels from different anatomical regions. The resulting skeletons were rasterized using a Bresenham algorithm [82] to create volumes of binary vessel masks. To simulate the change of contrast agent concentration in the vessels over time after the insertion of a bolus, we modeled the bolus as a Gaussian distribution moving through the vessel tree and weighted voxels accordingly. These vessel structures were then forward projected from different angles to simulate vessel-only (DSA) images. Our experiments demonstrate that this simple and hemodynamically incorrect model of contrast agent variation is sufficient to generate realistic vessel structures that improve the performance of the trained networks.

To generate synthetic X-ray angiography images that serve as input to the networks, we combined the vessel-only projections with mask images from clinical CT acquisitions. To this end, we forward projected CT measurements of cadavers acquired with a photon-counting research prototype. We then picked isocenters of the forward projections uniformly at random within the whole-body CT acquisitions, thus generating mask images including all anatomical regions. These mask images were then combined with the vessel-only projections to create a paired dataset of X-ray images and DSA-like images.

To evaluate whether training on our synthetic data improves the performance of deep learning-based methods for DSA, we trained two different networks (one with adversarial loss and one with standard pixelwise loss) on our synthetic data and compared their performance to

networks trained on clinical DSA data of the lower extremities. We then evaluated the networks on static clinical C-arm acquisitions of the lower extremities as well as on dynamic cardiac, pelvis, abdomen, and bolus chase acquisitions. We also conducted an ablation study to investigate the impact of the amount and diversity with respect to anatomical regions of synthetic data on the performance of the networks.



▲ **Figure 4.1.** X-ray image (first column), conventional DSA (second column), DSA from a network trained on clinical data (third column), and DSA from a network trained on synthetic data (last column) for different anatomical regions (from top to bottom: cardiac, pelvis, abdomen). Conventional DSA (second column) is computed via subtraction of the contrast-enhanced X-ray image (first column) from a mask image without contrast media and contains severe motion artifacts if patient motion is present. Predictions from a network trained on synthetic data (last column) contain fewer artifacts than from a network trained on clinical DSA acquisitions (third column). Figure adapted from [72].

4.1.3 Results

Upon a quantitative and qualitative evaluation on static acquisitions, we find that training on synthetic data is competitive with a training on clinical acquisitions. However, all networks remove or hallucinate vessels to some extent and are therefore inferior to conventional DSA on these data.

On acquisitions where patient motion is inevitable (cardiac, pelvis, abdomen), we find that networks trained on synthetic data outperform networks trained on clinical data (Fig. 4.1). This is likely due to the fact that these data pose an OOD problem for networks trained on clinical data, where patient (background) anatomy as well as vessel structures are different. Networks trained on synthetic data, on the other hand, have seen similar anatomical structures and a greater variety of vessel structures during training. As expected, conventional DSAs for these data contain severe motion artifacts (Fig. 4.1, second column).

4.2 Paper: Training of a Deep Learning-Based Digital Subtraction Angiography Method Using Synthetic Data

The following pages contain the full paper [72]. To improve readability, the supplementary material of this paper is provided in Appendix C.

DOI: 10.1002/mp.16973

RESEARCH ARTICLE

MEDICAL PHYSICS

Training of a deep learning based digital subtraction angiography method using synthetic data

Lizhen Duan[1,2,3] | **Elias Eulig**[1,4] | **Michael Knaup**[1] | **Ralf Adamus**[5] | **Michael Lell**[5] | **Marc Kachelrieß**[1,6]

[1]Division of X-Ray Imaging and Computed Tomography, German Cancer Research Center (DKFZ), Heidelberg, Germany

[2]School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences (UCAS), Beijing, China

[3]Key Laboratory of Optical Engineering, Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu, China

[4]Faculty of Physics and Astronomy, Heidelberg University, Heidelberg, Germany

[5]Department of Radiology, Neuroradiology and Nuclear Medicine, Klinikum Nürnberg, Paracelsus Medical University, Nürnberg, Germany

[6]Medical Faculty Heidelberg, Heidelberg University, Heidelberg, Germany

Correspondence
Marc Kachelrieß, Division of X-Ray Imaging and Computed Tomography, German Cancer Research Center (DKFZ), Heidelberg, Germany.
Email: marc.kachelriess@dkfz.de

Funding information
China Scholarship Council, Grant/Award Number: 202104910370; Deutsches Krebsforschungszentrum; Helmholtz International Graduate School for Cancer Research

Abstract

Background: Digital subtraction angiography (DSA) is a fluoroscopy method primarily used for the diagnosis of cardiovascular diseases (CVDs). Deep learning-based DSA (DDSA) is developed to extract DSA-like images directly from fluoroscopic images, which helps in saving dose while improving image quality. It can also be applied where C-arm or patient motion is present and conventional DSA cannot be applied. However, due to the lack of clinical training data and unavoidable artifacts in DSA targets, current DDSA models still cannot satisfactorily display specific structures, nor can they predict noise-free images.
Purpose: In this study, we propose a strategy for producing abundant synthetic DSA image pairs in which synthetic DSA targets are free of typical artifacts and noise commonly found in conventional DSA targets for DDSA model training.
Methods: More than 7,000 forward-projected computed tomography (CT) images and more than 25,000 synthetic vascular projection images were employed to create contrast-enhanced fluoroscopic images and corresponding DSA images, which were utilized as DSA image pairs for training of the DDSA networks. The CT projection images and vascular projection images were generated from eight whole-body CT scans and 1,584 3D vascular skeletons, respectively. All vessel skeletons were generated with stochastic Lindenmayer systems. We trained DDSA models on this synthetic dataset and compared them to the trainings on a clinical DSA dataset, which contains nearly 4,000 fluoroscopic x-ray images obtained from different models of C-arms.
Results: We evaluated DDSA models on clinical fluoroscopic data of different anatomies, including the leg, abdomen, and heart. The results on leg data showed for different methods that training on synthetic data performed similarly and sometimes outperformed training on clinical data. The results on abdomen and cardiac data demonstrated that models trained on synthetic data were able to extract clearer DSA-like images than conventional DSA and models trained on clinical data. The models trained on synthetic data consistently outperformed their clinical data counterparts, achieving higher scores in the quantitative evaluation of peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) metrics for DDSA images, as well as accuracy, precision, and Dice scores for segmentation of the DDSA images.
Conclusions: We proposed an approach to train DDSA networks with synthetic DSA image pairs and extract DSA-like images from contrast-enhanced x-ray images directly. This is a potential tool to aid in diagnosis.

KEYWORDS
deep learning, digital subtraction angiography, fluoroscopy, synthetic training data

1 | INTRODUCTION

Digital subtraction angiography (DSA)[1,2] is a fluoroscopic subtraction technique,[3] primarily used for the diagnosis of cardiovascular diseases (CVDs) with high morbidity and mortality in adults,[4] such as arterial and venous occlusions and stenoses, coronary artery disease, cerebral thrombosis, and pulmonary embolism. A DSA exam is obtained by subtracting a mask image (an x-ray image acquired prior to the contrast agent injection) from all subsequent images, in which a contrast agent such as iodine alters the radiodensity of the vessels, leading to an image selectively showing the vessels without patient background. However, due to the subtraction step, this technique has one major disadvantage, its limitation to static data. Any motion of the C-arm or patient can prompt huge distortion and artifacts in the subtracted images,[2] which impair their clinical value.

To reduce the motion artifacts in DSA images, several image registration algorithms have been proposed, such as feature-based image registration,[5] intensity-based registration,[6] and using non-uniform Markov random field models.[7] Inspired by the remarkable performance achieved by deep learning in medical imaging,[8,9] several studies applying neural networks to generate DSA-like images from a single contrast-enhanced fluoroscopy image have been proposed. Eulig et al.[10] used deep convolutional neural networks to learn maskless DSA for static and dynamic acquisition protocols, called deep DSA (DDSA). Gao et al.[11] presented an adversarial training scheme to generate DSA-like image from single live image. Furthermore, DDSAs with different network structures were applied to lower extremity,[12] abdominal vasculature[13] and cerebral vascular imaging,[14,15] respectively. Compared to a conventional DSA, deep learning-based methods have two major advantages. On the one hand, these methods may potentially improve image quality and reduce the potential radiation dose by reducing image artifacts caused by misalignment in DSA. On the other hand, these methods can learn DSA-like images directly from fluoroscopic images without prior acquisition of mask images, which allows them to be applied to dynamic acquisition protocols, such as bolus injection chases.[10] However, all these methods require a large amount of clinical data for training, which are difficult to obtain and rarely publicly available. Furthermore, because all current DDSA models were trained on pairs of native fluoroscopic images and DSA images obtained using the naive conventional DSA method, the targets used for network training invariably contained various degrees of mask contribution. As a result, the predicted DSA-like images also contained various degrees of artifacts like those in DSA images, which is particularly evident in the application to abdominal and coronary data.[13-15]

In this paper, we present an approach to generate a large number of synthetic DSA image pairs by combining forward-projected computed tomography (CT) images with synthetic vascular projection images. The synthetic DSA image pairs are employed to train different convolutional neural networks (CNNs).[10,11,13,14] Compared to clinical DSA images, the synthetic DSA images are free of misalignment and noise artifacts. To validate the effectiveness and robustness of synthetic DSA dataset in this task, we also train the CNNs on a clinical DSA dataset. The experimental results on different patient exams show that models trained on synthetic data can predict more vessels and fewer artifacts than models trained on clinical data. The quantitative results in peak signal-to-noise ratio (PSNR), structural similarity index measure[16] (SSIM), accuracy, precision, and the Sørensen–Dice coefficient metrics also indicate the superiority of our synthetic data.

2 | METHODS

In this section, we describe how to generate synthetic DSA data and train CNN models after illustrating the drawbacks of clinical DSA image pairs for network training.

2.1 | Clinical DSA image pair

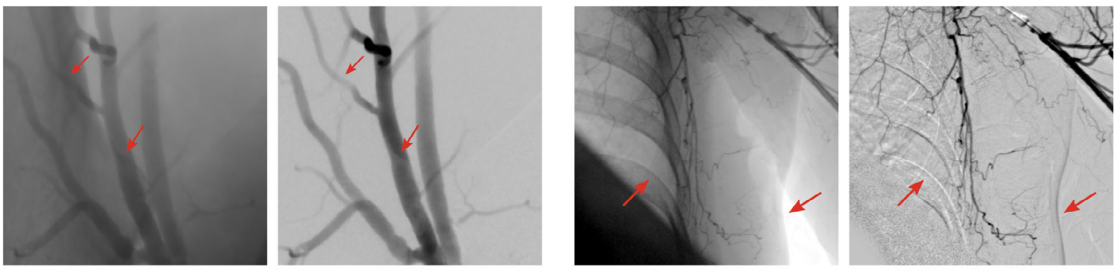
Assuming monochromatic x-ray radiation and no motion of the C-arm and patient, we can obtain an artifact-free DSA image. According to Lambert's law, the intensity of an x-ray image at time t is calculated with

$$I_t = I_0 e^{-\mu d_t} \tag{1}$$

where I_0 indicates the image before administration of contrast (the mask), μ is the attenuation coefficient of iodine at the desired x-ray energy, and d_t is the time-dependent iodine thickness concentration multiplied by the thickness of the contrast-enhanced vessels over the ray path. Then, an artifact-free DSA image is given as

$$f_{\text{afDSA}} = \ln(I_t) - \ln(I_0) = -\mu d_t \tag{2}$$

Unfortunately, we are typically not dealing with the situation in Equation (2). In practice, the x-ray images are acquired with polychromatic radiation and the detector signal is not identical to the intensity I. The vendor typically applies a nonlinear function to the measured signal, both implicitly in the detector when converting the x-ray



(a) Mask leakage (b) Motion artifact

FIGURE 1 (a) Left: x-ray image with contrast agent, right: conventional DSA image. Mask leakage: in the case of polychromatic x-ray radiation and $G \neq \ln$, a small contribution of the mask can be seen in the DSA image. The red arrows show two distinct brighter vessel blocks caused by the bone structure in the mask. (b) Left: x-ray image with contrast agent, right: conventional DSA image. Motion artifacts: in the case of misalignment due to motion, a large contribution (red arrows) of the mask can be seen in the DSA image.

photons into a digital signal, and explicitly when processing this signal. This processing may include, but is not limited to, applying a gain factor, converting the signal to an integer-valued signal, clipping signal values that are out of range, and applying a window function or response curve or, in general, a nonlinear mapping to the values before displaying them or before storing them in some formats. This mapping, here denoted as $G(\cdot)$, varies between acquisitions, organ programs and vendors. Most likely, at least for the purpose of generating DSA data, the vendor will choose G similar to ln function.

For standard DSA, $G(I_t)$ and $G(I_0)$ are used to get a DSA image. Let f_{cDSA} be the DSA image in clinical implementation, it is computed via

$$f_{\text{cDSA}} = G(I_t) - G(I_0) \approx -\mu d(t) \tag{3}$$

Generally, the native fluoroscopic image $G(I_t)$ and the DSA image f_{cDSA} are used as the input and target (clinical DSA image pair) for DDSA model training.[10,11,13,14] In most clinical cases, the contribution of the mask can be visible to varying degrees in the DSA image. Figure 1 illustrates this situation. This contribution may not be perceived by a human reader but may adversely affect the training of the network with the measurements.

2.2 | Synthetic DSA image pair

In the following, we describe how to generate synthetic DSA image pairs using clinical CT data and synthetic vessel skeletons.

2.2.1 | Forward-projected CT images

Mask images for the synthetic data were generated by forward-projecting clinical CT images acquired with a SOMATOM CounT photon counting CT research prototype (Siemens Healthineers, Forchheim, Germany). The cadaver measurements have been acquired in scope of a forensic study in close collaboration with the Institute of Forensic and Traffic Medicine (Prof. Sarah Heinze), Heidelberg University, Heidelberg, Germany, after being approved by the local ethics review board (S 388/2014). The CT data included the human head, torso as well as upper and lower extremities. The parameter settings of the projection simulation are listed in Table 1. We picked in total 105 isocenters within the eight whole-body CT scans uniformly at random, and generated 72 projections (with 5° increment) per isocenter, resulting in 7,560 forward-projections. Figure 2 shows several representative CT forward projections of different body regions.

TABLE 1 Parameter settings for CT projection images and vascular images.

Parameter	CT projection images	Vascular images
Source data size	1024 × 1024 × 1024 voxels	1024 × 1024 × 1024 voxels
Tube voltage	N.A. (water pre-correction applied)	70 keV
Focal-spot-to-isocenter distance	785 mm	785 mm
Isocenter-to-detector distance	415 mm	415 mm
Detector size	1024 × 1024 pixels	1024 × 1024 pixels
Horizontal pixel size	0.3 mm	0.3 mm
Vertical pixel size	0.3 mm	0.3 mm
Number of projections	72	16
Angular increment	5°	11.25°

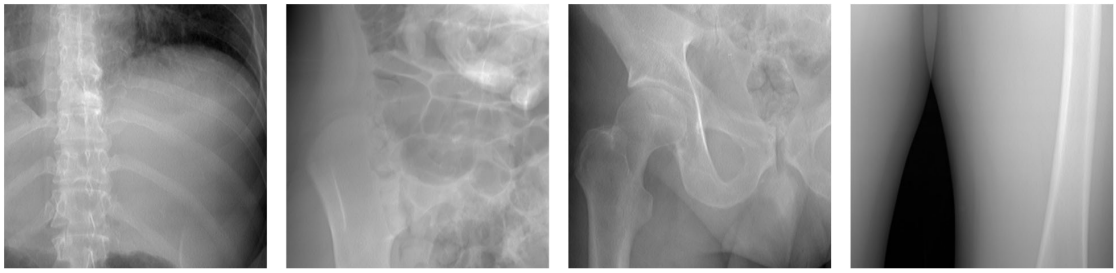


FIGURE 2 Representative CT forward projections of the clinical data used in this study.

TABLE 2 Parameter settings for simulating 3D vessel skeletons with stochastic L-systems.

Parameter	Values
Initial diameter d_0	$\{5x, 2 \leq x \leq 12, x \in \mathbb{N}^+\}$ pixels
Ratio of vessel length to diameter r_{ld}	$\{x, 4 \leq x \leq 9, x \in \mathbb{N}^+\}$
Iteration number n_{iter}	$\{x, 4 \leq x \leq 11, x \in \mathbb{N}^+\}$
Ratio of a branch vessel diameter to its parent vessel diameter r_{dd}	$\{1/\sqrt[3]{2}, 1/\sqrt[3]{1.5}, 1/\sqrt[3]{1.2}\}$

$\mathbb{N}^+$ denotes the set of positive integers.

2.2.2 | Simulate vascular projection images

We applied stochastic L-systems[17] to generate vessel skeletons. The highly flexible rules of the stochastic L-systems allow for a rich variety of generated vessels. A modified Bresenham algorithm[18] and diameter information were then used to generate vascular volumes in 3D. The volumes generated by this method had a pixel value 1 for vessels and 0 otherwise. To simulate the change of contrast agent concentration in vessels over time we applied a Gaussian function.

Given the distance $d(\boldsymbol{r})$ of the voxel $\boldsymbol{r}$ to the location of the bolus injection along the vascular tree and the flow velocity v, the time when the bolus reaches $\boldsymbol{r}$ is given as $t_0 = d(\boldsymbol{r})/v$. Thus, the voxel value is weighted with

$$w(\boldsymbol{r}, t) = e^{-\frac{1}{2}(t - t_0(\boldsymbol{r}))^2/\sigma^2} \tag{4}$$

to mimic the flow of the contrast agent. The parameter σ denotes the temporal width of the Gaussian bolus. One should note that this type of flow simulation is neither physically nor hemodynamically correct. Nonetheless, empirically, we find that the Gaussian model helps the network to learn how to cope with temporally varying contrast concentrations (Appendix B.1).

In this work, we simulated a total of 1,584 vessel skeletons with different vessel structures. Table 2 lists the parameter settings used to generate these 3D vessel skeletons. The starting diameter of the synthetic vessel was the given initial diameter plus a random value sampled from a Gaussian normal distribution with a mean of 0 pixel and a standard deviation of 5 pixels. All 3D vessel skeletons were rescaled to a cube of 1024 pixels on each side before the Gaussian function was applied to simulate contrast flow.

Finally, by projecting these vessel skeletons from 16 different angles, a total of 25,344 vascular projection images were obtained. Detailed information about this vascular projection simulation is listed in Table 1. For flow simulation, we set v to 50 pixels/s, σ to 10 s, and t to 12 s. Figure 3 shows some examples of simulated vascular projection images.

2.2.3 | Generate synthetic DSA image pairs

We generate an x-ray image after administration of contrast F_t from a CT forward projection image B and a vascular projection image P_t via

$$\begin{aligned} F_t &= F_0 + T_t \\ F_0 &= 1 - B \\ T_t &= -\alpha P_t \end{aligned} \tag{5}$$

where B and P_t are both normalized to [0, 1], and α is an empirical ratio of the maximum value of a clinical mask image and the maximum value of the corresponding DSA image. F_0 is regarded as a mask image, and F_t and T_t constitute a synthetic DSA image pair. The α was empirically set to a random value sampled uniformly from [0, 0.6] in our study. The α setting can be adjusted to suit different application data. To match noise levels to those of conventional fluoroscopic images, we introduced noise based on the statistics associated with photons into the forward projections and measured the noise levels in the images using signal-to-noise ratio (SNR). Additional details of this process can be found in Appendix C. Figure 4 shows two synthetic DSA image pairs generated with our method.

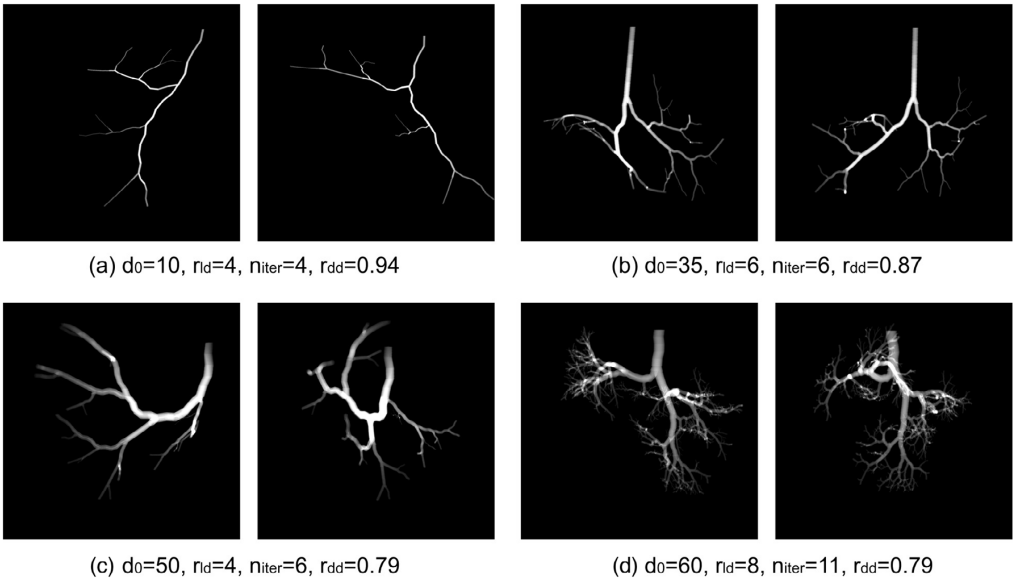


FIGURE 3 Simulated vascular projection images at different angles. Parameter settings for generating vessel skeletons: (a) $d_0 = 10, r_{ld} = 4, n_{iter} = 4, r_{dd} = 0.94$ (b) $d_0 = 35, r_{ld} = 6, n_{iter} = 6, r_{dd} = 0.87$ (c) $d_0 = 50, r_{ld} = 4, n_{iter} = 6, r_{dd} = 0.79$ (d) $d_0 = 60, r_{ld} = 8, n_{iter} = 11, r_{dd} = 0.79$.

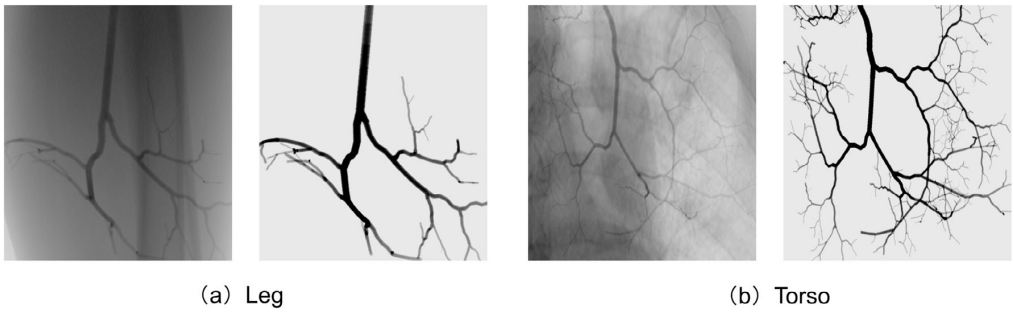


FIGURE 4 Synthetic DSA image pairs. (a) Left: synthetic x-ray image with a leg mask. Right: corresponding DSA image. (b) Left: synthetic x-ray image with a torso mask. Right: corresponding DSA image.

2.3 | Network and training

2.3.1 | Training data

In our experiments we compared training methods with two types of datasets, the synthetic data generated using our method, and a conventional clinical dataset as used for training in previous works.[12-15]

As described in Sections 2.2.1 and 2.2.2, a total of 7,560 CounT projection images and 25,344 vascular projection images were generated and utilized for DSA image pair generation. Thereafter, synthetic data were arbitrarily split into a training and test dataset on a per-isocenters and per-vessel basis with a approximately 9:1 ratio. A synthetic DSA image pair was generated by randomly picking a CT projection image and a vascular projection image using Equation (5).

As clinical training data 58 static C-arm acquisitions (3,400 images) of the human leg were used. For evaluation of the trained models (both on synthetic and clinical data), we used six additional C-arm acquisitions (544 images), three of which were dynamic bolus chase studies (cf. Figure 10).

To analyze the impact of the amount of clinical images in the synthetic dataset on network performance, additional experiments with different training set size of clinical image data were conducted and the results are presented in Appendix B.2. In the following, all CT data

are used to train the networks unless otherwise stated, since the main purpose of the paper is to introduce a new and effective method to generate synthetic data for DDSA model training, and this approach achieves the best results among all experiments.

Both synthetic and clinical data were normalized to zero-mean unit-variance before feeding them into the network.

2.3.2 | Networks

A 2D U-net[8] and a conditional geneartive adversarial network (GAN)[19] were employed as DDSA models. The 2D U-net[8] is utilized to learn a DSA image from an x-ray image directly. The conditional GAN consists of a generator for learning a DSA image from an x-ray image and a discriminator for judging whether an x-ray image and a DSA image is a ground truth DSA pair or generated. Detailed architectures of the networks are given in Appendix A.1.

2.3.3 | Implementation

The U-net model was trained by minimizing the pixel-wise mean absolute error (MAE) loss between the output and the ground truth. The conditional GAN model was trained with a pixelwise MAE loss and a modified adversarial loss.[20,21] Additional training details are provided in Appendix A.2. Our implementation was based on the PyTorch framework (version 1.0.0, 2018; Windows) and run on a workstation with 4 NVIDIA GeForce RTX 3090 GPUs, each with 24 GB of memory. With the same parameter settings, we trained both the U-net and the GAN on the synthetic dataset and the clinical dataset, respectively. In the following, we present the results for networks with minimal validation loss.

3 | EXPERIMENTAL RESULTS

We test models on three clinical datasets. All data in the three datasets come from a different patient. **(1) Leg data**: This dataset includes three static leg data from the clinical test dataset. The three data have an image size of 1024 × 1024 pixels and contain 36, 127, and 100 images, respectively. These DSA sequences contain few artifacts and can be used as ground truth. **(2) Abdomen and pelvis data**. This dataset includes two image sequences from two exams on the abdomen and pelvis. The two data have image sizes of 960 × 1240 and 1240 × 960 pixels, each containing 21 images. Due to the breathing motion, these DSA sequences contain motion artifacts which severely affects their clinical value. **(3) Cardiac data**. This dataset includes two image sequences from coronary exams. The two data have an image size of 1024 × 1024 pixels, and contain 47 and 25 images, respectively. Due to the heartbeat, these DSA sequences are also with significant motion artifacts. All three datasets are provided by Klinikum Nürnberg (Nürnberg, Germany).

In the following we show images computed from taking the minimum value of each pixel over all frames in a temporal sequence (minimum image). Either for a DSA image sequence or a predicted DSA-like image sequence, the minimum image shows all true vessels and wrongly inferred vessels of the method through the sequence.

Figure 5 presents results on the test leg dataset, including conventional DSA, output of U-net trained on clinical data with pixel-wise loss (DDSA-pix-clc), output of U-net trained on synthetic data with pixel-wise loss (DDSA-pix-syn), output of GAN trained on clinical DSA data with pixel-wise loss and adversarial loss (DDSA-adv-clc), and output of GAN trained on synthetic data with pixel-wise loss and adversarial loss (DDSA-adv-syn). For different DDSA models that training on synthetic data can produce clearer DSA-like images than training on the clinical DSA data. Results of DDSA-pix-syn contain fewer wrongly inferred vessels and more detailed vessels than DDSA-pix-clc. Both DDSA-adv-clc and DDSA-adv-syn incorrectly inferred and lost more vessels than DDSA-pix-clc and DDSA-pix-syn. In total, tests on legs data show that models trained on synthetic data are competitive with models trained on clinical data.

Figures 6 and 7 present results on the abdomen and pelvis and cardiac datasets. Compared to conventional DSA images, results of DDSA-pix-syn and DDSA-adv-syn contain clearer vessels and much less motion artifacts and noise. Results of DDSA-pix-syn and DDSA-adv-syn contain more detailed vessels and fewer artifacts than those of DDSA-pix-clc and DDSA-adv-clc, respectively. Tests on abdomen and cardiac data show that models trained on synthetic data outperform conventional DSA method and models trained on clinical data.

To quantitatively evaluate the performance of different methods, metrics of PSNR and SSIM were adopted. Table 3 lists the PSNR and SSIM scores between results produced by different models and DSA images on the leg dataset. It can be concluded that the models trained on our synthetic data perform better than those trained on clinical data.

To further assess the similarity of vessels in DDSA images to DSA images, we segmented the DDSA results using Otsu thresholding algorithm on the leg dataset. We performed segmentation on gamma-enhanced DDSA images and followed it with erosion to achieve more accurate segmentation. As shown in Figure 8, the segmentations produced by models trained on synthetic data exhibit more distinct vessel features than those produced by models trained on clinical data. We

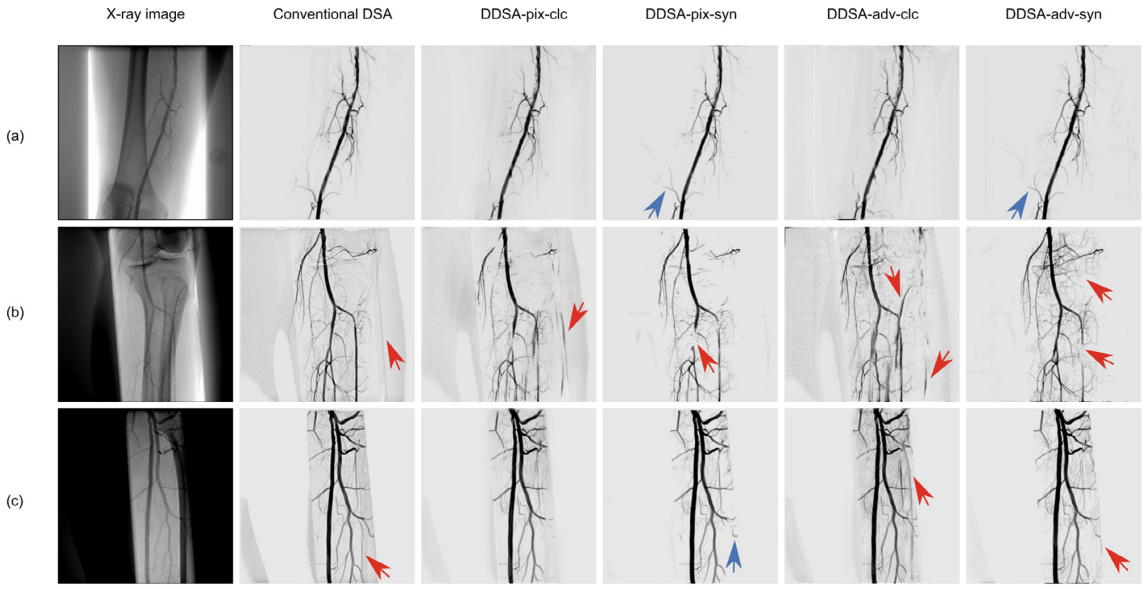


FIGURE 5 Results on clinical leg data: (a)–(c) represent different test data. Left to right: the minimum images of original fluoro sequences, conventional DSA sequences, results of DDSA-pix-clc, DDSA-pix-syn, DDSA-adv-clc, and DDSA-adv-syn. Blue arrows indicate more vessels extracted by DDSA models trained on synthetic data. Red arrows indicate motion artifacts produced by conventional DSA method, and vessels wrongly inferred and lost by DDSA models.

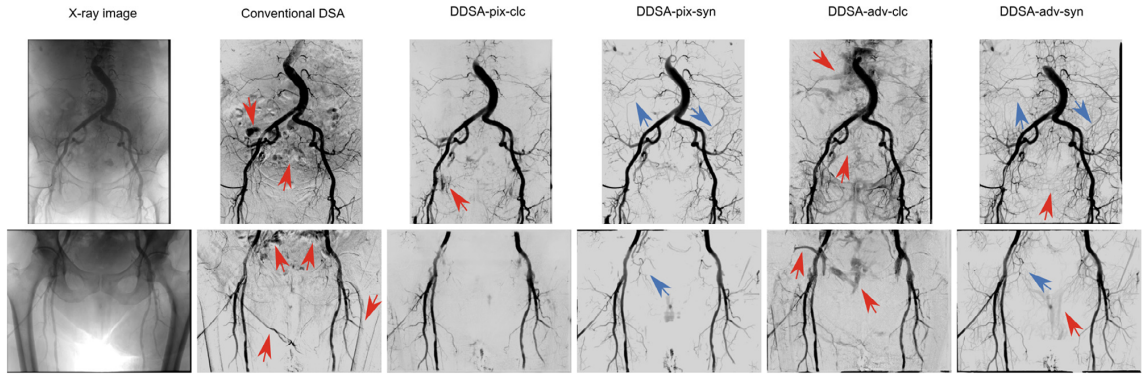


FIGURE 6 Results on clinical abdomen and pelvis data. Left to right: the minimum images of original fluoro sequences, conventional DSA sequences, results of DDSA-pix-clc, results of DDSA-pix-syn, results of DDSA-adv-clc, and results of DDSA-adv-syn. Blue arrows indicate more or clearer vessels extracted by DDSA models trained on synthetic data. Red arrows indicate motion artifacts produced by conventional DSA method, and wrongly inferred vessels by DDSA models.

TABLE 3 Quantitative comparison on the clinical leg dataset: PSNR and SSIM with 95% confidence intervals.

	PSNR			SSIM		
Method	**Figure 5a**	**Figure 5b**	**Figure 5c**	**Figure 5a**	**Figure 5b**	**Figure 5c**
DDSA-pix-clc	42.50 ± 1.97	34.86 ± 1.11	35.61 ± 3.30	0.9980 ± 0.0009	0.9780 ± 0.0018	0.9791 ± 0.0193
DDSA-pix-syn	**43.75 ± 1.49**	**36.32 ± 0.90**	35.89 ± 1.62	**0.9982 ± 0.0008**	**0.9820 ± 0.0026**	**0.9845 ± 0.0013**
DDSA-adv-clc	40.16 ± 1.24	32.80 ± 0.41	33.09 ± 4.23	0.9962 ± 0.0014	0.9645 ± 0.0026	0.9684 ± 0.0276
DDSA-adv-syn	**42.94 ± 1.33**	**35.29 ± 1.39**	**34.24 ± 3.10**	**0.9978 ± 0.0010**	**0.9787 ± 0.0029**	**0.9811 ± 0.0013**

Note: The bold values denote the statistical significance of results of training on synthetic data compared to the results of training on clinical data, as determined by a Wilcoxon signed-rank test with a significance level $\alpha = 0.05$.

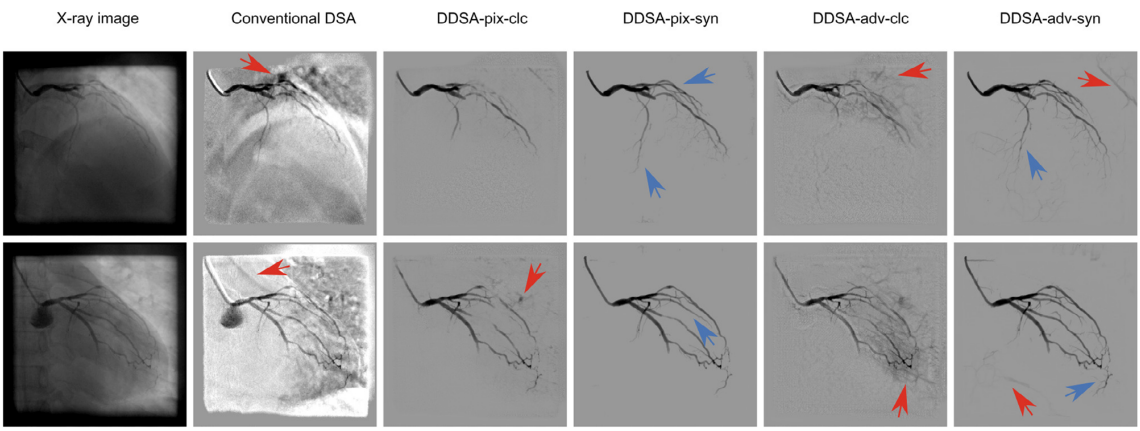


FIGURE 7 Results on clinical cardiac data. Left to right: original fluoro image, conventional DSA, results of DDSA-pix-clc, results of DDSA-pix-syn, results of DDSA-adv-clc, and results of DDSA-adv-syn. Blue arrows indicate more and clearer vessels extracted by DDSA models trained on synthetic data. Red arrows indicate motion artifacts and wrongly inferred vessels by DDSA models.

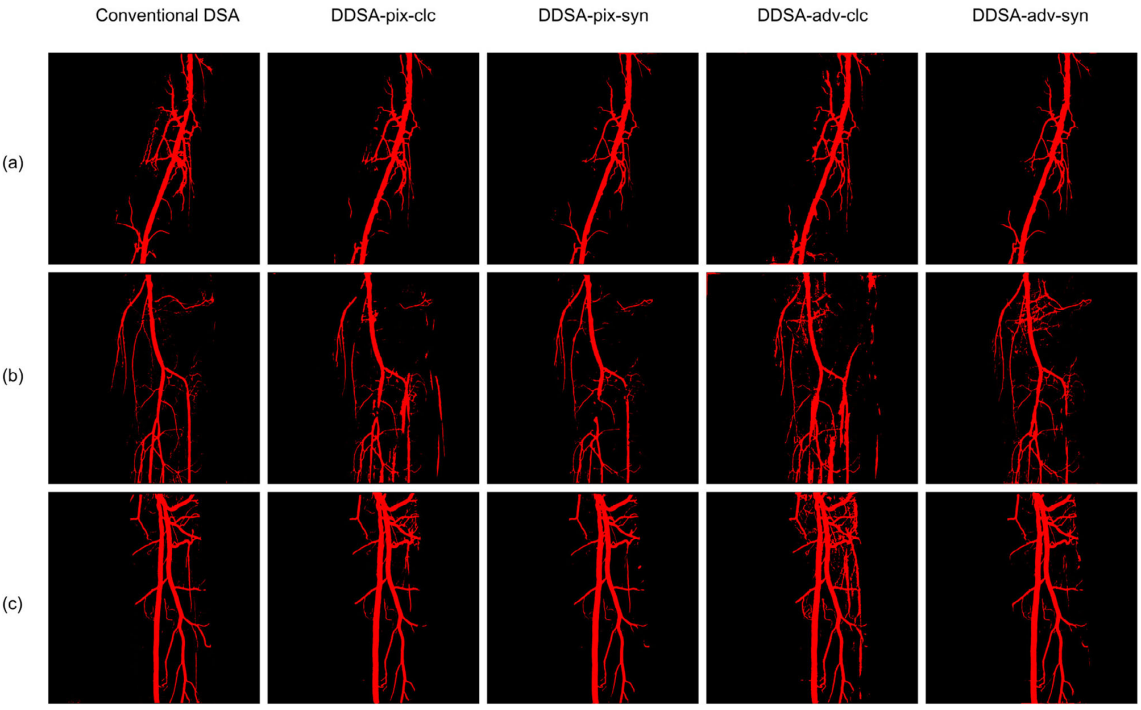


FIGURE 8 Segmentation results on clinical leg data: (a)–(c) represent different test data. Left to right: the segments correspond to the minimum images of conventional DSA sequences, results of DDSA-pix-clc, results of DDSA-pix-syn, results of DDSA-adv-clc, and results of DDSA-adv-syn.

quantitatively evaluate the performance of different DDSA segments by comparing them to the DSA segmentations, which serve as the ground truth, using accuracy, precision, and Dice score. Pixels containing vessels are considered as positive and vice versa. The evaluation results are presented in Table 4. According to these evaluation metrics, the segmentations of DDSA-pix-syn and DDSA-adv-syn exhibit higher quality compared to those of DDSA-pix-clc and DDSA-adv-clc, respectively. The improvement in these metrics indicates that the models trained on our synthetic data exhibit superior overall accuracy and more accurate vessel identification.

TABLE 4 Quantitative comparison on the segmentation results of clinical leg dataset: accuracy, precision, and Dice score with 95% confidence intervals.

Method	Accuracy			Precision			Dice score		
	Figure 8a	Figure 8b	Figure 8c	Figure 8a	Figure 8b	Figure 8c	Figure 8a	Figure 8b	Figure 8c
DDSA-pix-clc	0.989 ± 0.003	0.973 ± 0.004	0.980 ± 0.008	0.869 ± 0.321	0.770 ± 0.050	0.677 ± 0.361	0.807 ± 0.391	0.604 ± 0.262	0.719 ± 0.216
DDSA-pix-syn	0.988 ± 0.004	**0.977 ± 0.006**	**0.983 ± 0.007**	**0.898 ± 0.165**	**0.938 ± 0.037**	**0.757 ± 0.007**	0.822 ± 0.124	**0.638 ± 0.244**	0.735 ± 0.199
DDSA-adv-clc	0.985 ± 0.006	0.963 ± 0.005	0.973 ± 0.015	0.801 ± 0.157	0.596 ± 0.056	0.596 ± 0.482	0.779 ± 0.190	0.504 ± 0.228	0.658 ± 0.340
DDSA-adv-syn	**0.989 ± 0.004**	**0.974 ± 0.005**	**0.979 ± 0.010**	**0.887 ± 0.004**	**0.817 ± 0.027**	**0.722 ± 0.342**	**0.835 ± 0.124**	**0.622 ± 0.243**	**0.670 ± 0.247**

Note: The bold values denote the statistical significance of results of training on synthetic data compared to the results of training on clinical data, as determined by a Wilcoxon signed-rank test with a significance level $\alpha = 0.05$.

To delve deeper into the factors contributing to the enhanced performance of the network model trained on synthetic data, we conducted additional experiments on the U-net model using three newly generated synthetic datasets. These datasets were structured as follows: Dataset A included an equal number of vascular images and CT projection images of the leg, in analogy to the clinical dataset; Dataset B comprised an equal number of CT projection images of the leg, similar to the clinical dataset, but incorporating all generated vascular images; and Dataset C contains an equal number of vascular images as the clinical dataset, but includes all CT projection images originating from different anatomical regions. It is worth noting that we only had CT projection images of the legs from four patients, whereas the clinical DSA training data were derived from 58 patients. This apparent discrepancy makes the comparison between Dataset A and the clinical data not entirely fair.

The models trained on the three aforementioned synthetic datasets using pixel-wise loss, were further evaluated by testing them on the three clinical datasets. We designate the models trained on datasets A, B, and C as DDSA-equ-all, DDSA-equ-masks, and DDSA-equ-vessels, respectively.

The results are presented in Figure 9. The quantitative results on the test leg dataset are shown in Table 5. The networks trained using different synthetic datasets effectively eliminate background artifacts. DDSA-equ-all incorrectly identifies some bone structures as vessels. With more simulated vessels used for training, DDSA-equ-masks effectively eliminates these misidentifications on the leg dataset. With more CT images from different anatomical regions used for training, DDSA-equ-masks effectively corrected most of these misidentifications across the three anatomical datasets. However, DDSA-equ-masks failed to identify specific large-sized vessels in Figure 9c and f. This discrepancy can be attributed to the smaller number of vessels used in generating simulated data compared to CT projection images, possibly leading to overfitting on leg data. This overfitting causes the network to misclassify large vessels as bones, resulting in their omission. The models trained on synthetic datasets predicted more complete vessels on both abdominal and cardiac data compared to the models trained on clinical data. The quantitative results on the leg data as shown in Tables 3 and 5 demonstrate that the results of DDSA-equ-masks are better than those of DDSA-pix-clc and close to those of DDSA-pix-syn. This suggests that when employing an equivalent number of anatomical images and regions as the clinical data, our synthesis method can incorporate more vascular images during training, thereby improving the quality of DSA images generated by the network.

We perform additional experiments by evaluating networks trained on both synthetic and clinical data on a

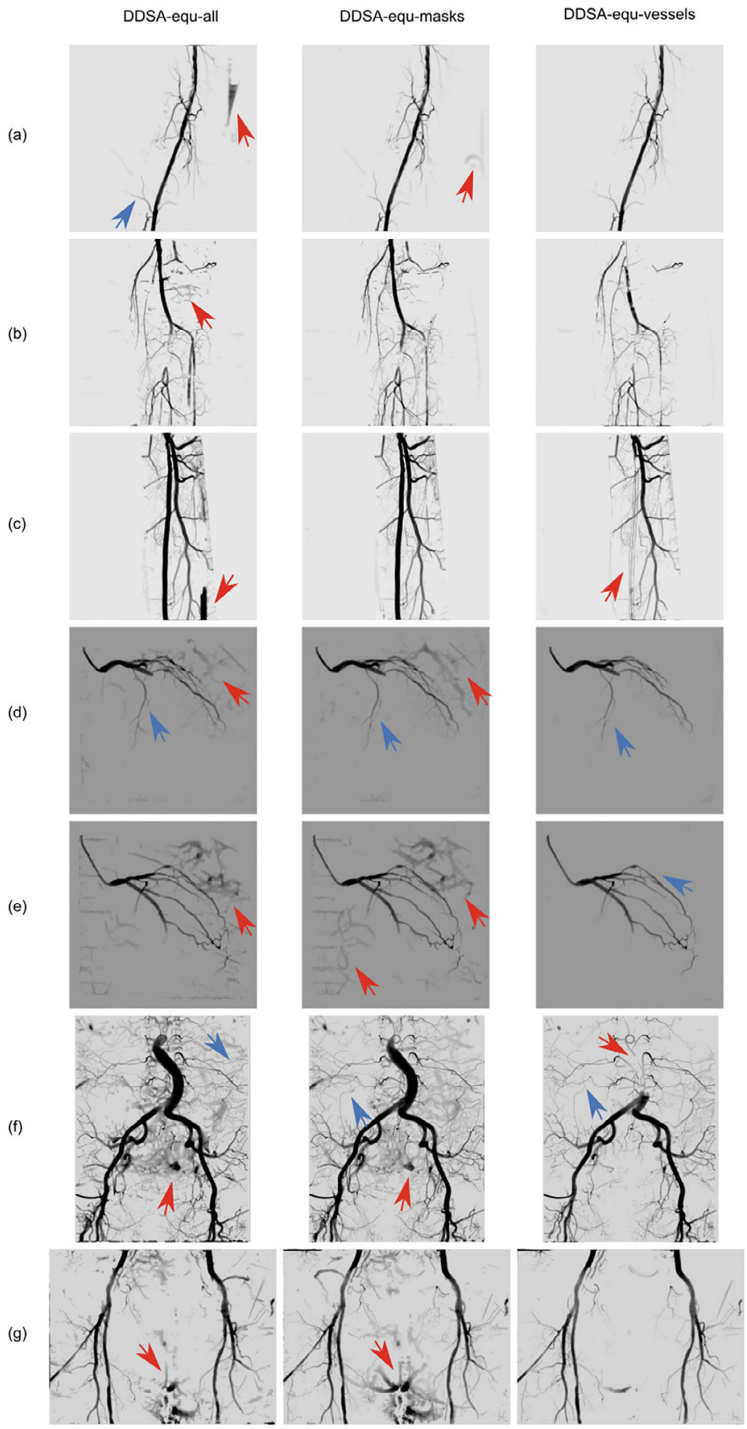


FIGURE 9 Results of training with different numbers of anatomical regions and images: (a)–(g) represent different test data. Left to right: results of DDSA-equ-all, results of DDSA-equ-masks, and results of DDSA-equ-vessels. Blue arrows indicate more and clearer vessels extracted by DDSA models trained on synthetic data. Red arrows indicate wrongly inferred vessels by DDSA models.

dynamic bolus chase study. In this study, the patient table was manually maneuvered to track the movement of the contrast agent as it traveled through the femoral and tibial vessels. Due to the motion of C-arm, conventional DSA is not feasible for this exam. Therefore, a direct comparison between DDSA and conventional DSA cannot be made. Figure 10 exhibits a frame captured from the femur, where models trained on synthetic data can extract similar DSA images as models trained on clinical data. However, there are some fine vascular structures

TABLE 5 Quantitative comparison of networks trained with different simulated datasets on the clinical leg dataset: PSNR and SSIM with 95% confidence intervals.

Method	PSNR			SSIM		
	Figure 9a	Figure 9b	Figure 9c	Figure 9a	Figure 9b	Figure 9c
DDSA-equ-all	40.42±0.31	34.79±0.65	32.67±2.20	0.9977±0.0007	0.9797±0.0025	0.9808±0.0011
DDSA-equ-masks	43.53±1.35	36.27±0.66	35.94±2.18	0.9981±0.0008	0.9818±0.0022	0.9840±0.0009
DDSA-equ-vessels	43.81±1.74	35.57±0.71	32.13±6.53	0.9982±0.0009	0.9788±0.0030	0.9812±0.0053

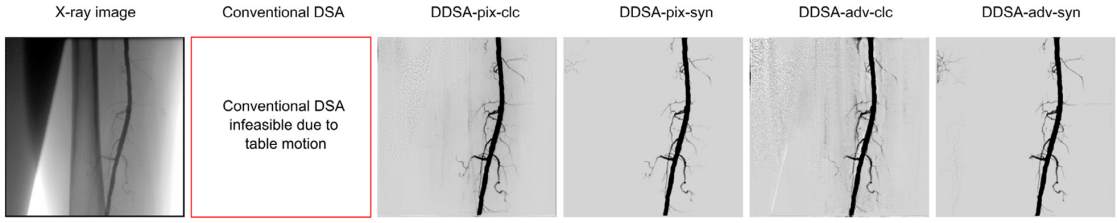


FIGURE 10 Results on a bolus chase study. Left to right: original fluoro image, conventional DSA infeasible due to table motion, results of DDSA-pix-clc, results of DDSA-pix-syn, results of DDSA-adv-clc, and results of DDSA-adv-syn.

that are not captured by the model trained on synthetic data. This could be attributed to the limited variety of sources for CT projection images. Although there were 7,560 CT projection images used to generate the synthetic DSA image pairs, these images were derived exclusively from CT scans of eight patients. Notably, only four of the patients' scans contained anatomical regions of the legs.

4 | DISCUSSION

In this paper, we proposed a method to generate abundant synthetic DSA image pairs for DDSA model training. To our knowledge, this is the first proposal to generate DSA data with synthetic vessels and CT data and use them to train DDSA models.

We simulated a total of 1,584 3D vessel skeletons with different sizes, lengths and branches using stochastic L-systems. These simulated vessel skeletons and 105 CT data sets are used for DSA image pairs generation. In order to test the reliability and validity of the synthetic data for DDSA models training, we trained a U-net and a GAN, on synthetic data and on clinical data, respectively. The test results on the leg, abdomen and pelvis and heart datasets show that networks trained on synthetic DSA data can extract visually clearer DSA-like images than networks trained on clinical DSA data. Our results contain much less artifacts and noise. Of particular note is that our models perform better than conventional DSAs on abdomen and cardiac data, as can be seen in Figures 6 and 7. The quantitative evaluation of PSNR and SSIM for DSA images, along with accuracy, precision, and Dice for DSA segmented images, consistently shows the superiority of our synthetic data.

Since the models trained on our synthetic data drastically suppress the contribution of background, it can be inferred that our methods are superior to the existing abdominal vascular DDSA study,[13] even though we tested on different exams in Figure 6. All these superiorities of our synthetic DSA dataset are mainly due to the rich artifact-free synthetic DSA images. The synthetic method enables the incorporation of different anatomical images, thus improving the ability of the network to accurately predict DSA images in various anatomical regions. This work can be extended to other vessel extraction studies.[22,23]

As indicated by the blue arrows in Figures 5, 6, and 7, DDSA-adv-clc and DDSA-adv-syn omit and wrongly infer more vessels than DDSA-pix-clc and DDSA-pix-syn, respectively. Some of the wrongly inferred structures are so similar to actual vessels that it is difficult to visually identify them as incorrectly inferred vessels, and thus we recommend using the pixel-loss trained network instead of the adversarial trained network for this task.

The limitation of this work is that the DSA images predicted by our methods still contain certain artifacts coming from the mask images and may omit certain vessels in particular data. These wrongly inferred and omitted vessels occur primarily at the boundaries of bones and tissues and where blood vessels and bone are in close proximity. This is likely due to the fact that in the presented method mask images and synthetic vascular projection images are picked independently. This problem may be alleviated by improving the way CT projection images and synthetic vascular projection images are combined, such as generating more DSA

data in which the vascular branches are close to and almost parallel to bones.

5 | CONCLUSIONS

This study developed a method generating synthetic DSA image pairs to train neural network models for DSA image extraction with CT images and simulated vessels. Benefiting from diverse synthetic training data and accurate synthetic DSA targets, models trained on the synthetic data outperform models trained on clinical data in both visual and quantitative assessments. This approach compensates for the paucity and inadequacy of clinical DSA data. Its application can also be expanded to include cerebral and retinal angiography.

ACKNOWLEDGMENTS

Parts of the reconstruction and simulation software were provided by RayConStruct GmbH, Nürnberg, Germany. Author Lizhen Duan would like to acknowledge the financial support received from the China Scholarship Council (CSC), grant number 202104910370. Elias Eulig was supported in part by the Helmholtz International Graduate School for Cancer Research, Heidelberg, Germany.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts to disclose.

DATA AVAILABILITY STATEMENT

Code and data are available at https://github.com/D-I-Z/synthetic-ddsa.

REFERENCES

1. Chilcote WA, Modic MT, Pavlicek WA, et al. Digital subtraction angiography of the carotid arteries: a comparative study in 100 patients. *Radiology*. 1981;139:287-295.
2. Brody WR. Digital subtraction angiography. *IEEE Trans Nucl Sci*. 1982;29:1176-1180.
3. Hanafee W, Stout P. Subtraction technic. *Radiology*. 1962;79:658-661.
4. Kralj V, Brkić Biloš I. Morbidity and mortality from cardiovascular diseases. *Cardiol Croat*. 2013;8:373-378.
5. Zhang X, Zhang F, Li R. DSA image registration based on 3D space-time detection. *Procedia Eng*. 2010;7:426-431.
6. Hipwell JH, Penney GP, McLaughlin RA, et al. Intensity-based 2-D-3-D registration of cerebral angiograms. *IEEE Trans Med Imaging*. 2003;22:1417-1426.
7. Sundarapandian M, Kalpathi R, Manason VD. DSA image registration using non-uniform MRF model and pivotal control points. *Comput Med Imaging Graph*. 2013;37:323-336.
8. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer; 2015:234-241.
9. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60-88.
10. Eulig E, Maier J, Knaup M, Koenig T, Hörndler K, Kachelrieß M. Deep DSA (DDSA): learning mask-free digital subtraction angiography for static and dynamic acquisitions protocols using a deep convolutional neural network. In: *ECR 2019 Book of Abstracts, Insights into Imaging 10 (Suppl. 1)*. 2019:S379.
11. Gao Y, Song Y, Yin X, et al. Deep learning-based digital subtraction angiography image generation. *Int J Comput Assist Radiol Surg*. 2019;14:1775-1784.
12. Eulig E, Maier J, Knaup M, Koenig T, Hörndler K, Kachelrieß M. Learned digital subtraction angiography (Deep DSA): method and application to lower extremities. In: *15th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine*. Vol 11072. SPIE; 2019:360-363.
13. Yonezawa H, Ueda D, Yamamoto A, et al. Maskless 2-dimensional digital subtraction angiography generation model for abdominal vasculature using deep learning. *J Vasc Interv Radiol*. 2022;33:845-851.
14. Ueda D, Katayama Y, Yamamoto A, et al. Deep learning–based angiogram generation model for cerebral angiography without misregistration artifacts. *Radiology*. 2021;299:675-681.
15. Vepa A, Choi A, Nakhaei N, et al. Weakly-supervised convolutional neural networks for vessel segmentation in cerebral angiography. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022:585-594.
16. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*. 2004;13:600-612.
17. Galarreta-Valverde MA, Macedo MM, Mekkaoui C, Jackowski MP. Three-dimensional synthetic blood vessel generation using stochastic L-systems. In: *Medical Imaging 2013: Image Processing*. Vol 8669. SPIE; 2013:414-419.
18. Bresenham JE. Algorithm for computer control of a digital plotter. *IBM Syst J*. 1965;4:25-30.
19. Gui J, Sun Z, Wen Y, et al. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Trans Knowl Data Eng*. 2021;35:3313-3332.
20. Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: *International Conference on Machine Learning*. PMLR; 2017:214-223.
21. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of Wasserstein Gans. *Adv Neural Inf Process Syst*. 2017;30:5769-5779.
22. Lamy J, Merveille O, Kerautret B, Passat N. A benchmark framework for multiregion analysis of vesselness filters. *IEEE Trans Med Imaging*. 2022;41:3649-3662.
23. Qin B, Mao H, Liu Y, et al. Robust PCA unrolling network for super-resolution vessel extraction in x-ray coronary angiography. *IEEE Trans Med Imaging*. 2022;41:3087-3098.
24. Huang L, Qin J, Zhou Y, et al. Normalization techniques in training dnns: Methodology, analysis and application. *IEEE Trans Pattern Anal Mach Intell*. 2023;45:10173-10196.
25. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. In: *International Conference on Learning Representations, ICLR;* 2015:1-13.

How to cite this article: Duan L, Eulig E, Knaup M, Adamus R, Lell M, Kachelrieß M. Training of a deep learning based digital subtraction angiography method using synthetic data. *Med Phys*. 2024;51:4793–4810. https://doi.org/10.1002/mp.16973

Chapter 5

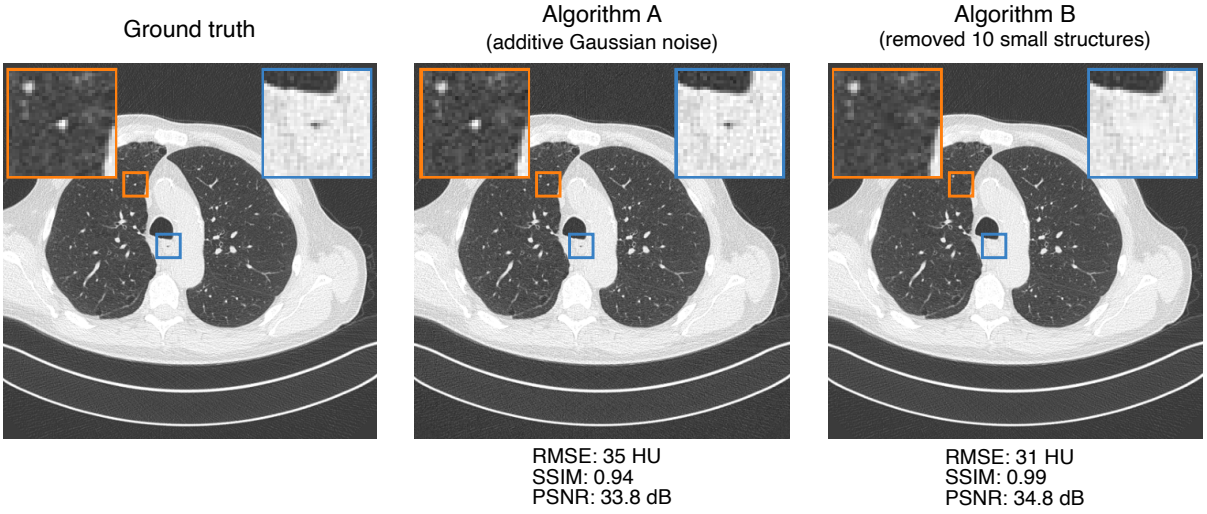
Discussion & Outlook

We now want to summarize and discuss the results of the three presented projects and their broader implications for the field of CT and medical imaging. In particular, we want to focus on how our work addressed the three challenges with DNNs for medical imaging discussed in the introduction (Chapter 1): scarcity of data, interpretability and robustness, and fair evaluation of models. We also want to highlight some limitations of our work and suggest directions for future research to overcome the challenges of deep learning-based methods in medical imaging.

5.1 Benchmarking Low-Dose CT Denoising Networks

In our first project presented in Chapter 2, we proposed a benchmark framework for assessing deep learning-based LDCT denoising algorithms in a fair and reproducible manner. We identified several flaws in the experimental setup of many of these studies that limit the comparability and reproducibility of the results and their claimed improvements. To overcome these flaws, we introduced a unified dataset, a set of eight popular deep learning-based LDCT denoising algorithms, a rigorous hyperparameter optimization, and a set of suitable evaluation metrics. Using our benchmark suite, we evaluated the performance of the eight algorithms and found that most of them showed only marginal improvements over previous methods, with many performing similarly to RED-CNN [37], one of the earliest networks proposed for this task. GAN-based models showed some superiority in preserving radiomic features, particularly in high-noise settings like low-dose chest scans. However, the newer methods did not consistently outperform older techniques. The findings suggest a need for more rigorous validation and evaluation in LDCT denoising research, with the proposed benchmark providing a crucial foundation for future developments.

We want to emphasize that similar "reality checks" have been performed in other areas of deep learning research such as language modelling [83] and metric learning [84], often deriving similar conclusions: Once hyperparameter optimization and other experimental details are controlled for, newer methods often do not outperform older methods. Many of the issues that were



▲ **Figure 5.1.** A toy example to illustrate the limitations of traditional IQA metrics for evaluating medical image reconstruction algorithms. The ground truth image (left) is altered by two hypothetical reconstruction algorithms. Algorithm A adds a small amount of Gaussian noise ($\mu = 0\,\text{HU}$, $\sigma = 35\,\text{HU}$) to the image, while Algorithm B removes ten small structures (see zoom-ins for two examples thereof). While Algorithm A is likely to be less harmful in a clinical setting, traditional image metrics commonly used to evaluate deep learning-based medical image reconstruction algorithms such as RMSE, SSIM, and PSNR would favor Algorithm B. $C = -600\,\text{HU}$, $W = 1500\,\text{HU}$.

analyzed in these and our work transfer to other areas of deep learning in medical imaging. In particular, the lack of unified datasets and open-source code is a common issue in many areas of medical imaging research, and we hope that our work will inspire researchers to release their code and data to allow for better reproducibility and comparability of results in deep learning-based medical imaging. In the following, we want to highlight some limitations of this work and suggest directions for future research.

We hope that researchers will use our framework in the future to evaluate and benchmark their novel algorithms. However, since many of the newly proposed methods are dual-domain, meaning they employ networks that operate in both raw-data and image domain, it would be beneficial to extend our benchmark suite from image-domain to dual-domain. Similarly, for simplicity, we only considered 2D networks in our benchmark suite. However, many of the newer methods consider multiple axial slices or even 3D volumes, and it would be interesting to extend our benchmark suite to support a fair evaluation of 3D networks as well.

While LDCT image denoising is one of the most prevalent applications of deep learning for CT imaging, there are many other research areas in medical imaging that would benefit from a similar benchmark such as fast magnetic resonance imaging (MRI) [85], sparse-view CT [86], MAR [54], or positron emission tomography (PET) image reconstruction [87]. We hope that our work will inspire researchers in these areas to develop similar benchmark suites to evaluate and compare their algorithms.

We acknowledge that a clinically meaningful evaluation of deep-learning based algorithms for

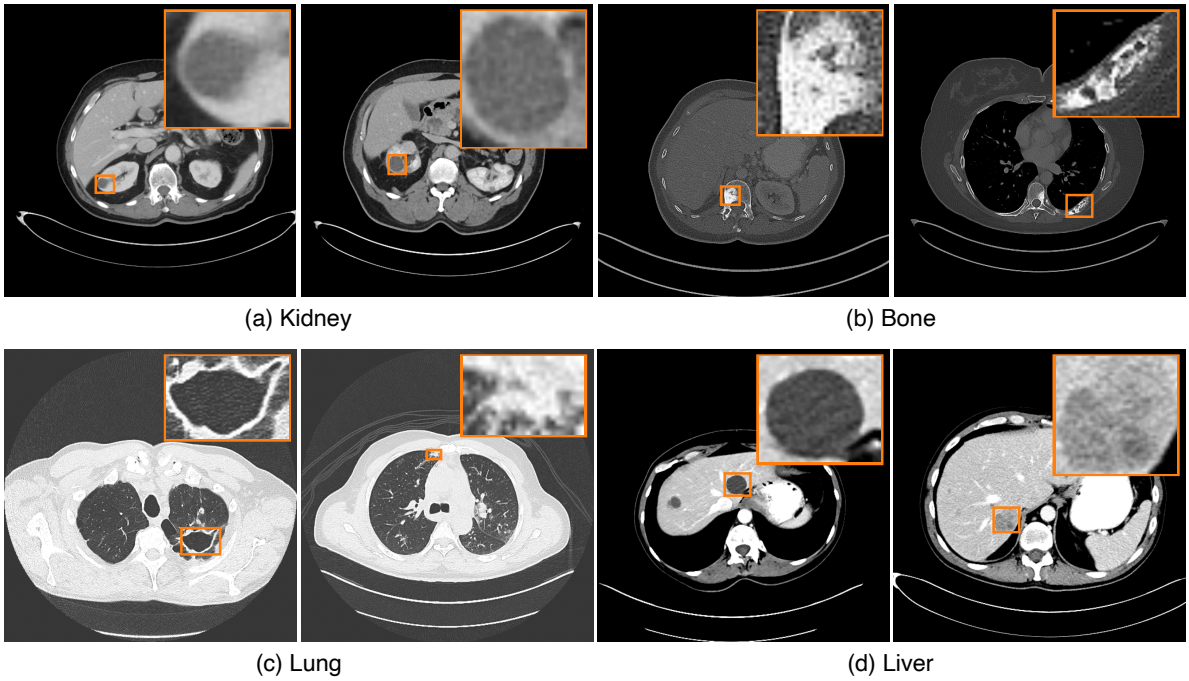
medical imaging remains an open challenge. This work made some contributions towards better IQA metrics for medical images by utilizing metrics that have previously been shown to better correlate with human readers and by introducing a novel metric to evaluate the preservation of radiomic features in the denoised images. Nonetheless, more work has to be invested in the development of better metrics, particularly those that are sensitive to alterations of small structures or lesions. Common full-reference IQA metrics such as SSIM or RMSE often fail in this regard because small features in image space contribute only little to the overall measure (see Fig. 5.1 for an illustrative example of this problem). If segmentations of *clinically relevant* structures were available, this could be alleviated by overrepresenting these structures, similar to the related problem of class imbalances in image segmentation and object recognition tasks [88]. However, the definition of what is clinically relevant relies on the definition of a single or multiple *clinical tasks* which is often difficult in practice. If such tasks can be defined, IQA could also be performed indirectly by evaluating the performance on one or multiple downstream clinical tasks (e.g., tumor segmentation or detection) using human readers or ML-based proxies thereof. Evaluation using such proxies is particularly relevant given the increased use of ML for computer-aided diagnosis (CAD), where reconstructed medical image data must be suitable for human readers and ML algorithms simultaneously.

5.2 Invariances of Low-Dose CT Denoising Networks

In this project, presented in Chapter 3, we investigated the invariances of deep learning-based low-dose CT image denoising networks. We developed a framework that reconstructs invariances by adapting a method previously used for image classification networks to the task of LDCT denoising. Applying this framework to four popular LDCT denoising networks, we found that all networks are predominantly invariant to noise level and noise realization, as expected. We further proposed two methods to analyze sampled invariances in terms of their alteration of anatomical structures. Using these methods we found that the networks are also invariant to anatomical structures, to some extent. Such invariances could lead to the removal or alteration of clinically relevant features, highlighting the need for careful analysis of model behavior in medical imaging applications. This work has some limitations and gives rise to future research directions which we want to discuss below.

Similar to our benchmarking project (see Sec. 5.1), while LDCT denoising is a prevalent application of deep learning in medical imaging, it would be interesting to analyze the invariances of networks in other medical imaging tasks, including other modalities such as MRI or PET. Furthermore, such invariances could also be analyzed in the context of medical image analysis tasks such as segmentation, detection, or classification of pathologies.

In this work, we analyzed the sampled invariances only regarding their alteration of anatomical structures and did not consider whether these alterations are also clinically relevant. This could be done using a human reader study, where radiologists are asked to detect pathologies in both the original images and the sampled invariance images. Alternatively, a more automated



▲ **Figure 5.2.** Four exemplar pathologies of different organs contained in the open *DeepLesion* dataset [89, 90]. Zoom-ins correspond to bounding-box annotations of the respective lesions. In total, the dataset comprises roughly 32k annotated CT slices, containing a variety of pathologies. Training a lesion classification network on such data could help to analyze the invariances of deep learning-based medical imaging algorithms regarding their alteration of pathological features. (a) $C = 50\,\mathrm{HU}$, $W = 400\,\mathrm{HU}$; (b) $C = 400\,\mathrm{HU}$, $W = 1800\,\mathrm{HU}$; (c) $C = -600\,\mathrm{HU}$, $W = 1500\,\mathrm{HU}$; (d) $C = 50\,\mathrm{HU}$, $W = 200\,\mathrm{HU}$.

approach could be to first train a classifier on open datasets containing a variety of annotated pathologies [89, 90] which could then be used to classify both types of images (Fig. 5.2). In either case, the results of the human reader study or the classifier can be used to determine whether the networks are invariant to pathologies in the images. Related to this is the question of the effect of different datasets on the type of invariances learned by the networks. In particular, we hypothesize that the amount or type of pathologies that a network has seen during training also influences the invariances learned by the network. Further investigation on this topic could be done by training networks on datasets with different amounts or types of pathologies and analyzing the invariances of these networks.

Lastly, it would be interesting to connect invariances to the related concept of *hallucinations*. Hallucinations is an umbrella term for predictions made by a DNN that are highly plausible but factually wrong. The term is particularly used in the context of large language models (LLMs), which have been shown to hallucinate various types of information [91]. In medical imaging, hallucinations could be the addition or removal of clinically incorrect features to an image [92, 93, 94, 95]. It seems plausible that networks that are invariant to specific features in the input images are also more likely to hallucinate them and investigating this relationship would be an interesting direction for future research.

5.3 Synthetic Training Data for Deep Learning-Based Digital Subtraction Angiography

In the third project presented in Chapter 4, we tackled the scarcity of data for DSA by simulating paired data for training DNNs for synthetic DSA. We did so by combining vessels that were simulated using stochastic L-systems with forward projections of clinical CT data. We showed that training on these data is superior to training on real data as it allows the networks to better generalize to anatomical regions for which the acquisition of static DSA data is infeasible. We verified that this improvement is architecture-agnostic and translates to networks trained using both pixelwise and adversarial loss functions. There are some limitations and future directions for this work that we will discuss below.

While some form of artificiality in the training data is common for most tasks in medical imaging (c.f. Chapter 1, *Scarcity of data*), the use of synthetic data for training DNNs in medical imaging also comes with its own set of challenges. In particular, for most simulated data it remains unknown how well the models generalize to real data. Given that the real-world forward model is often unknown or too complex to be modeled accurately, some discrepancy between the synthetic and real data is to be expected. On its own, this would not be a problem if the discrepancy could be quantified. However, with clinical paired data not being available for many applications, quantitative evaluation of the generalization of the models to real data is difficult or impossible. To circumvent this issue, many researchers acquire paired data via phantom measurements, where scans can be repeated at different dose levels and different artifacts can be avoided by removing metal implants, performing slit scans, or using other techniques. While this is suitable for many classical algorithms which are linear and shift-invariant, it is not for deep-learning based methods. Here, the evaluation on phantom data poses an OOD problem since the algorithms are usually only trained on clinical data or simulations thereof. Even for methods that are trained on a mixture of phantom and patient data (GE Healthcare's TrueFidelity™[96] is a prominent example), assessments of performance on phantom measurements may not be representative of the performance on clinical data.

Furthermore, similar to our discussion of IQA metrics in Sec. 5.1, even if a ground truth is available (as is the case for static DSA), quantitative evaluation is limited by the lack of suitable metrics. Standard image metrics, such as RMSE, weigh all pixels in the image equally, and thus small alterations of vessel structures (e.g., addition or removal of stenoses) are not well captured (Fig. 5.1). In our work, we addressed this issue by performing an additional quantitative evaluation using some common image segmentation metrics, namely precision, recall, and the Dice-Sørensen coefficient (DSC). To this end, we segmented vessels in both the ground truth DSA and the synthetic DSA via simple thresholding and compared the segmentations using the aforementioned metrics. Future work could leverage human annotations of vessel masks and pathological features to extend this evaluation to dynamic DSA data and to evaluate whether the networks are able to reconstruct clinically relevant features.

Chapter 6

Summary

In this thesis, three projects were presented that address challenges associated with deep learning-based medical imaging. The first project introduced a benchmarking framework to enable fair evaluation of low-dose CT denoising algorithms. The second project investigated these networks in terms of their invariances to features in the input images. Finally, the third project tackled the problem of data scarcity for digital subtraction angiography by leveraging synthetic training data for training deep learning models for synthetic DSA.

Note that the projects discussed in this thesis also highlight opportunities for deep learning in medical imaging in general and X-ray imaging and CT in particular. In the context of LDCT denoising, DNNs have demonstrated superior performance compared to traditional methods and will serve as a key component in future efforts to reduce patient doses further. In other areas of medical imaging, deep learning enabled entirely new methodologies and procedures that were previously unimaginable. Examples include risk-based tube current modulation [48, 49], image reconstruction from extremely undersampled data [97, 98], and the here discussed deep learning-based DSA.

If the remaining challenges are tackled and novel applications are discovered, deep learning-based methods for medical imaging can support clinicians, reduce costs, and ultimately improve patient care.

Bibliography

[1] C. H. McCollough and P. S. Rajiah. “Milestones in CT: Past, Present, and Future”. In: *Radiology* 309.1 (2023), e230803.

[2] W. C. Röntgen. “Ueber Eine Neue Art von Strahlen (Vorläufige Mittheilung)”. In: *Sitzungsberichte der Physikalisch-Medizinischen Gesellschaft zu Würzburg* (1895).

[3] National Health Service (NHS). *Diagnostic Imaging Dataset Statistical Release*. Tech. rep. 1. 2023.

[4] United Nations Scientific Committee on the Effects of Atomic Radiation. *Sources, Effects and Risks of Ionizing Radiation, UNSCEAR 2012 Report: Report to the General Assembly, with Scientific Annexes A and B*. United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) Reports. UN, 2015.

[5] A. M. Cormack. “Representation of a Function by Its Line Integrals, with Some Radiological Applications”. In: *Journal of Applied Physics* 34.9 (1963), pp. 2722–2727.

[6] G. N. Hounsfield. “Method of and Apparatus for Examining a Body by Radiation Such as X or Gamma Radiation”. US3919552A. 1975.

[7] S. J. Swensen, J. R. Jett, J. A. Sloan, D. E. Midthun, T. E. Hartman, A.-M. Sykes, G. L. Aughenbaugh, F. E. Zink, S. L. Hillman, G. R. Noetzel, R. S. Marks, A. C. Clayton, and P. C. Pairolero. “Screening for Lung Cancer with Low-Dose Spiral Computed Tomography”. In: *American Journal of Respiratory and Critical Care Medicine* 165.4 (2002), pp. 508–513.

[8] P. T. Johnson, D. G. Heath, L. V. Hofmann, K. M. Horton, and E. K. Fishman. “Multidetector-Row Computed Tomography with Three-Dimensional Volume Rendering of Pancreatic Cancer: A Complete Preoperative Staging Tool Using Computed Tomography Angiography and Volume-Rendered Cholangiopancreatography”. In: *Journal of Computer Assisted Tomography* 27.3 (2003-05/2003-06), p. 347.

[9] H. Takagi, R. Tanaka, K. Nagata, R. Ninomiya, K. Arakita, J. D. Schuijf, and K. Yoshioka. “Diagnostic Performance of Coronary CT Angiography with Ultra-High-Resolution CT: Comparison with Invasive Coronary Angiography”. In: *European Journal of Radiology* 101 (2018), pp. 30–37.

[10] J. T. P. D. Hallinan, C. H. Tan, and U. Pua. “Emergency Computed Tomography for Acute Pelvic Trauma: Where Is the Bleeder?” In: *Clinical Radiology* 69.5 (2014), pp. 529–537.

[11] V. Tacher, A. Radaelli, M. Lin, and J.-F. Geschwind. “How I Do It: Cone-Beam CT during Transarterial Chemoembolization for Liver Cancer”. In: *Radiology* 274.2 (2015), pp. 320–334.

[12] L. Pung, M. Ahmad, K. Mueller, J. Rosenberg, C. Stave, G. L. Hwang, R. Shah, and N. Kothary. “The Role of Cone-Beam CT in Transcatheter Arterial Chemoembolization for Hepatocellular Carcinoma: A Systematic Review and Meta-analysis”. In: *Journal of vascular and interventional radiology: JVIR* 28.3 (2017), pp. 334–341.

[13] D.-J. Kim, I. Chul-Nam, S.-E. Park, D.-R. Kim, J.-S. Lee, B.-S. Kim, G.-M. Choi, J. Kim, and J.-H. Won. “Added Value of Cone-Beam Computed Tomography for Detecting Hepatocellular Carcinomas and Feeding Arteries during Transcatheter Arterial Chemoembolization Focusing on Radiation Exposure”. In: *Medicina* 59.6 (2023), p. 1121.

[14] J. D. Louie, N. Kothary, W. T. Kuo, G. L. Hwang, L. V. Hofmann, M. L. Goris, A. H. Iagaru, and D. Y. Sze. “Incorporating Cone-beam CT into the Treatment Planning for Yttrium-90 Radioembolization”. In: *Journal of Vascular and Interventional Radiology* 20.5 (2009), pp. 606–613.

[15] M. Rodríguez-Fraile, A. Ezponda, F. Grisanti, V. Morán, M. Calvo, P. Berián, A. M. de la Cuesta, L. Sancho, M. Iñarrairaegui, B. Sangro, and J. I. Bilbao. “The Joint Use of 99mTc-MAA-SPECT/CT and Cone-Beam CT Optimizes Radioembolization Planning”. In: *EJNMMI Research* 11.1 (2021), pp. 1–11.

[16] G. Landry and C.-h. Hua. “Current State and Future Applications of Radiological Image Guidance for Particle Therapy”. In: *Medical Physics* 45.11 (2018), e1086–e1095.

[17] D. J. Brenner, R. Doll, D. T. Goodhead, E. J. Hall, C. E. Land, J. B. Little, J. H. Lubin, D. L. Preston, R. J. Preston, J. S. Puskin, E. Ron, R. K. Sachs, J. M. Samet, R. B. Setlow, and M. Zaider. “Cancer Risks Attributable to Low Doses of Ionizing Radiation: Assessing What We Really Know”. In: *Proceedings of the National Academy of Sciences of the United States of America* 100.24 (2003), p. 13761.

[18] D. I. Portess, G. Bauer, M. A. Hill, and P. O’Neill. “Low-Dose Irradiation of Nontransformed Cells Stimulates the Selective Removal of Precancerous Cells via Intercellular Induction of Apoptosis”. In: *Cancer Research* 67.3 (2007), pp. 1246–1253.

[19] M. Tubiana, L. E. Feinendegen, C. Yang, and J. M. Kaminski. “The Linear No-Threshold Relationship Is Inconsistent with Radiation Biologic and Experimental Data”. In: *Radiology* 251.1 (2009), pp. 13–22.

[20] L. Mullenders, M. Atkinson, H. Paretzke, L. Sabatier, and S. Bouffler. "Assessing Cancer Risks of Low-Dose Radiation". In: *Nature Reviews Cancer* 9.8 (2009), pp. 596–604.

[21] Y. F. Ali, F. A. Cucinotta, L. Ning-Ang, and G. Zhou. "Cancer Risk of Low Dose Ionizing Radiation". In: *Frontiers in Physics* 8 (2020).

[22] M. K. Kalra, M. M. Maher, T. L. Toth, L. M. Hamberg, M. A. Blake, J.-A. Shepard, and S. Saini. "Strategies for CT Radiation Dose Optimization". In: *Radiology* 230.3 (2004), pp. 619–628.

[23] D. J. Brenner and E. J. Hall. "Computed Tomography–an Increasing Source of Radiation Exposure". In: *The New England Journal of Medicine* 357.22 (2007), pp. 2277–2284.

[24] R. Singh, S. R. Digumarthy, V. V. Muse, A. R. Kambadakone, M. A. Blake, A. Tabari, Y. Hoi, N. Akino, E. Angel, R. Madan, and M. K. Kalra. "Image Quality and Lesion Detection on Deep Learning Reconstruction and Iterative Reconstruction of Submillisievert Chest and Abdominal CT". In: *American Journal of Roentgenology* 214.3 (2020), pp. 566–573.

[25] A. Ries, T. Dorosti, J. Thalhammer, D. Sasse, A. Sauter, F. Meurer, A. Benne, T. Lasser, F. Pfeiffer, F. Schaff, and D. Pfeiffer. "Improving Image Quality of Sparse-View Lung Tumor CT Images with U-Net". In: *European Radiology Experimental* 8 (2024), p. 54.

[26] K. Lange and R. Carson. "EM Reconstruction Algorithms for Emission and Transmission Tomography". In: *Journal of Computer Assisted Tomography* 8.2 (1984), pp. 306–316.

[27] C. Kamphuis and F. Beekman. "Accelerated Iterative Transmission CT Reconstruction Using an Ordered Subsets Convex Algorithm". In: *IEEE Transactions on Medical Imaging* 17.6 (1998), pp. 1101–1105.

[28] H. Erdogan and J. A. Fessler. "Ordered Subsets Algorithms for Transmission Tomography". In: *Physics in Medicine & Biology* 44.11 (1999), p. 2835.

[29] A. Ziegler, T. Köhler, and R. Proksa. "Noise and Resolution in Images Reconstructed with FBP and OSC Algorithms for CT". In: *Medical Physics* 34.2 (2007), pp. 585–598.

[30] Z. Yu, J.-B. Thibault, C. A. Bouman, K. D. Sauer, and J. Hsieh. "Fast Model-Based X-Ray CT Reconstruction Using Spatially Nonhomogeneous ICD Optimization". In: *IEEE Transactions on Image Processing* 20.1 (2011), pp. 161–175.

[31] M. Li, H. Yang, and H. Kudo. "An Accurate Iterative Reconstruction Algorithm for Sparse Objects: Application to 3D Blood Vessel Reconstruction from a Limited Number of Projections". In: *Physics in Medicine & Biology* 47.15 (2002), p. 2599.

[32] E. Y. Sidky, C.-M. Kao, and X. Pan. "Accurate Image Reconstruction from Few-Views and Limited-Angle Data in Divergent-Beam CT". In: *Journal of X-Ray Science and Technology* 14 (2006), pp. 119–139.

[33] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser. "Deep Convolutional Neural Network for Inverse Problems in Imaging". In: *IEEE Transactions on Image Processing* 26.9 (2017), pp. 4509–4522.

[34] J. Maier, E. Eulig, T. Vöth, M. Knaup, J. Kuntz, S. Sawall, and M. Kachelrieß. "Real-Time Scatter Estimation for Medical CT Using the Deep Scatter Estimation: Method and Robustness Analysis with Respect to Different Anatomies, Dose Levels, Tube Voltages, and Data Truncation". In: *Medical Physics* 46.1 (2019), pp. 238–249.

[35] P. Roser, A. Birkhold, A. Preuhs, C. Syben, L. Felsner, E. Hoppe, N. Strobel, M. Kowarschik, R. Fahrig, and A. Maier. "X-Ray Scatter Estimation Using Deep Splines". In: *IEEE Transactions on Medical Imaging* 40.9 (2021), pp. 2272–2283.

[36] H. Shan, A. Padole, F. Homayounieh, U. Kruger, R. D. Khera, C. Nitiwarangkul, M. K. Kalra, and G. Wang. "Competitive Performance of a Modularized Deep Neural Network Compared to Commercial Algorithms for Low-Dose CT Image Reconstruction". In: *Nature Machine Intelligence* 1.6 (2019), pp. 269–276.

[37] 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". In: *IEEE Transactions on Medical Imaging* 36.12 (2017), pp. 2524–2535.

[38] Z. Huang, J. Zhang, Y. Zhang, and H. Shan. "DU-GAN: Generative Adversarial Networks with Dual-Domain U-Net-based Discriminators for Low-Dose CT Denoising". In: *IEEE Transactions on Instrumentation and Measurement* 71 (2022), pp. 1–12.

[39] Z. Hong, D. Zeng, X. Tao, and J. Ma. "Learning CT Projection Denoising from Adjacent Views". In: *Medical Physics* 50.3 (2023), pp. 1367–1377.

[40] L. Yang, Z. Li, R. Ge, J. Zhao, H. Si, and D. Zhang. "Low-Dose CT Denoising via Sinogram Inner-Structure Transformer". In: *IEEE Transactions on Medical Imaging* 42.4 (2023), pp. 910–921.

[41] T. Würfl, F. C. Ghesu, V. Christlein, and A. Maier. "Deep Learning Computed Tomography". In: *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016*. Ed. by S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 432–440.

[42] T. Würfl, M. Hoffmann, V. Christlein, K. Breininger, Y. Huang, M. Unberath, and A. K. Maier. "Deep Learning Computed Tomography: Learning Projection-Domain Weights From Image Domain in Limited Angle Problems". In: *IEEE Transactions on Medical Imaging* 37.6 (2018), pp. 1454–1463.

[43] J. Adler and O. Öktem. "Learned Primal-Dual Reconstruction". In: *IEEE Transactions on Medical Imaging* 37.6 (2018), pp. 1322–1332.

[44] X. Han, J. Hong, M. Reyngold, C. Crane, J. Cuaron, C. Hajj, J. Mann, M. Zinovoy, H. Greer, E. Yorke, G. Mageras, and M. Niethammer. "Deep-Learning-Based Image Registration and Automatic Segmentation of Organs-at-Risk in Cone-Beam CT Scans from High-Dose Radiation Treatment of Pancreatic Cancer". In: *Medical Physics* 48.6 (2021), pp. 3084–3095.

[45] J. Maier, E. Eulig, S. Dorn, S. Sawall, and M. Kachelrieß. "Real-Time Patient-Specific CT Dose Estimation Using a Deep Convolutional Neural Network". In: *2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC)*. 2018, pp. 1–3.

[46] M. S. Lee, D. Hwang, J. H. Kim, and J. S. Lee. "Deep-Dose: A Voxel Dose Estimation Method Using Deep Convolutional Neural Network for Personalized Internal Dosimetry". In: *Scientific Reports* 9.1 (2019), p. 10308.

[47] T. I. Götz, C. Schmidkonz, S. Chen, S. Al-Baddai, T. Kuwert, and E. W. Lang. "A Deep Learning Approach to Radiation Dose Estimation". In: *Physics in Medicine & Biology* 65.3 (2020), p. 035007.

[48] L. Klein, C. Liu, J. Steidel, L. Enzmann, M. Knaup, S. Sawall, A. Maier, M. Lell, J. Maier, and M. Kachelrieß. "Patient-Specific Radiation Risk-Based Tube Current Modulation for Diagnostic CT". In: *Medical Physics* 49.7 (2022), pp. 4391–4403.

[49] C. Liu, L. Klein, Y. Huang, E. Baader, M. Lell, M. Kachelrieß, and A. Maier. "Two-View Topogram-Based Anatomy-Guided CT Reconstruction for Prospective Risk Minimization". In: *Scientific Reports* 14.1 (2024), p. 9373.

[50] A. M. Vepa, Z. Yang, A. Choi, J. Joo, F. Scalzo, and Y. Sun. "Integrating Deep Metric Learning with Coreset for Active Learning in 3D Segmentation". In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024.

[51] L. Yu, M. Shiung, D. Jondal, and C. H. McCollough. "Development and Validation of a Practical Lower-Dose-Simulation Tool for Optimizing Computed Tomography Scan Protocols". In: *Journal of Computer Assisted Tomography* 36.4 (2012), p. 477.

[52] S. E. Divel and N. J. Pelc. "Accurate Image Domain Noise Insertion in CT Images". In: *IEEE Transactions on Medical Imaging* 39.6 (2020), pp. 1906–1916.

[53] Y. Lyu, W.-A. Lin, H. Liao, J. Lu, and S. K. Zhou. "Encoding Metal Mask Projection for Metal Artifact Reduction in Computed Tomography". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Ed. by A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz. Cham: Springer International Publishing, 2020, pp. 147–157.

[54] B. Zhou, X. Chen, H. Xie, S. K. Zhou, J. S. Duncan, and C. Liu. "DuDoUFNet: Dual-domain under-to-Fully-Complete Progressive Restoration Network for Simultaneous Metal

Artifact Reduction and Low-Dose CT Reconstruction". In: *IEEE Transactions on Medical Imaging* 41.12 (2022), pp. 3587–3599.

[55] Z. C. Lipton. "The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability Is Both Important and Slippery." In: *Queue* 16.3 (2018), pp. 31–57.

[56] D. Hoiem, Y. Chodpathumwan, and Q. Dai. "Diagnosing Error in Object Detectors". In: *Computer Vision – ECCV 2012*. Ed. by A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid. Berlin, Heidelberg: Springer, 2012, pp. 340–353.

[57] F. R. Verdun, D. Racine, J. G. Ott, M. J. Tapiovaara, P. Toroi, F. O. Bochud, W. J. H. Veldkamp, A. Schegerer, R. W. Bouwman, I. H. Giron, N. W. Marshall, and S. Edyvean. "Image Quality in CT: From Physical Measurements to Model Observers". In: *Physica Medica* 31.8 (2015), pp. 823–843.

[58] G. P. Renieblas, A. T. Nogués, A. M. G. M.d, N. G. León, and E. G. del Castillo. "Structural Similarity Index Family for Image Quality Assessment in Radiological Images". In: *Journal of Medical Imaging* 4.3 (2017), p. 035501.

[59] K. Ohashi, Y. Nagatani, M. Yoshigoe, K. Iwai, K. Tsuchiya, A. Hino, Y. Kida, A. Yamazaki, and T. Ishida. "Applicability Evaluation of Full-Reference Image Quality Assessment Methods for Computed Tomography Images". In: *Journal of Imaging Informatics in Medicine* 36.6 (2023), pp. 2623–2634.

[60] A. Mason, J. Rioux, S. E. Clarke, A. Costa, M. Schmidt, V. Keough, T. Huynh, and S. Beyea. "Comparison of Objective Image Quality Metrics to Expert Radiologists' Scoring of Diagnostic Quality of MR Images". In: *IEEE Transactions on Medical Imaging* 39.4 (2020), pp. 1064–1072.

[61] W. Lee, F. Wagner, A. Galdran, Y. Shi, W. Xia, G. Wang, X. Mou, M. A. Ahamed, A. A. Z. Imran, J. E. Oh, K. Kim, J. T. Baek, D. Lee, B. Hong, P. Tempelman, D. Lyu, A. Kuiper, L. van Blokland, M. B. Calisto, S. Hsieh, M. Han, J. Baek, A. Maier, A. Wang, G. E. Gold, and J.-H. Choi. "Low-Dose Computed Tomography Perceptual Image Quality Assessment". In: *Medical Image Analysis* 99 (2025), p. 103343.

[62] L. S. Chow and R. Paramesran. "Review of Medical Image Quality Assessment". In: *Biomedical Signal Processing and Control* 27 (2016), pp. 145–154.

[63] E. Eulig, B. Ommer, and M. Kachelrieß. "Benchmarking Deep Learning-Based Low-Dose CT Image Denoising Algorithms". In: *Medical Physics* 51.12 (2024), pp. 8776–8788.

[64] P. Shunhavanich, S. S. Hsieh, and N. J. Pelc. "Fluid-Filled Dynamic Bowtie Filter: Description and Comparison with Other Modulators". In: *Medical Physics* 46.1 (2019), pp. 127–139.

[65] R. Fahrig, R. Dixon, T. Payne, R. L. Morin, A. Ganguly, and N. Strobel. "Dose and Image Quality for a Cone-Beam C-arm CT System". In: *Medical Physics* 33.12 (2006), pp. 4541–4550.

[66] I. Kim, H. Kang, H. J. Yoon, B. M. Chung, and N.-Y. Shin. "Deep Learning–Based Image Reconstruction for Brain CT: Improved Image Quality Compared with Adaptive Statistical Iterative Reconstruction-Veo (ASIR-V)". In: *Neuroradiology* 63.6 (2021), pp. 905–912.

[67] H. Sheikh and A. Bovik. "Image Information and Visual Quality". In: *IEEE Transactions on Image Processing* 15.2 (2006), pp. 430–444.

[68] C. McCollough, B. Chen, D. R. Holmes III, X. Duan, Z. Yu, L. Yu, S. Leng, and J. Fletcher. *Low Dose CT Image and Projection Data*. 2020.

[69] E. Eulig, F. Jäger, J. Maier, B. Ommer, and M. Kachelrieß. "Reconstructing and Analyzing the Invariances of Low-Dose CT Image Denoising Networks". In: *Medical Physics* 52.1 (2025), pp. 188–200.

[70] R. Rombach, P. Esser, and B. Ommer. "Making Sense of CNNs: Interpreting Deep Representations & Their Invariances with INNs". In: *European Conference on Computer Vision (ECCV)*. 2020, p. 18.

[71] E. Eulig, B. Ommer, and M. Kachelrieß. "Reconstructing Invariances of CT Image Denoising Networks Using Invertible Neural Networks". In: *7th International Conference on Image Formation in X-Ray Computed Tomography*. Vol. 12304. SPIE, 2022, pp. 169–173.

[72] L. Duan, E. Eulig, M. Knaup, R. Adamus, M. Lell, and M. Kachelrieß. "Training of a Deep Learning Based Digital Subtraction Angiography Method Using Synthetic Data". In: *Medical Physics* 51.7 (2024), pp. 4793–4810.

[73] T. G. Walker. "Acute Limb Ischemia". In: *Techniques in Vascular and Interventional Radiology*. "On Call" Emergencies for the Interventional Radiologist 12.2 (2009), pp. 117–129.

[74] S. Shaban, B. Huasen, A. Haridas, M. Killingsworth, J. Worthington, P. Jabbour, and S. M. M. Bhaskar. "Digital Subtraction Angiography in Cerebrovascular Disease: Current Practice and Perspectives on Diagnosis, Acute Treatment and Prognosis". In: *Acta Neurologica Belgica* 122.3 (2022), pp. 763–780.

[75] A. H. Sam and J. T. Teo. "Nephrology". In: *Rapid Medicine*. John Wiley & Sons, Ltd, 2018, pp. 121–134.

[76] E. Eulig, J. Maier, M. Knaup, T. Koenig, K. Hörndler, and M. Kachelrieß. "Learned Digital Subtraction Angiography (Deep DSA): Method and Application to Lower Extremities". In: *15th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine*. Vol. 11072. SPIE, 2019, pp. 360–363.

[77] Y. Gao, Y. Song, X. Yin, W. Wu, L. Zhang, Y. Chen, and W. Shi. "Deep Learning-Based Digital Subtraction Angiography Image Generation". In: *International Journal of Computer Assisted Radiology and Surgery* 14.10 (2019), pp. 1775–1784.

[78] D. Ueda, Y. Katayama, A. Yamamoto, T. Ichinose, H. Arima, Y. Watanabe, S. L. Walston, H. Tatekawa, H. Takita, T. Honjo, A. Shimazaki, D. Kabata, T. Ichida, T. Goto, and Y. Miki. "Deep Learning–Based Angiogram Generation Model for Cerebral Angiography without Misregistration Artifacts". In: *Radiology* 299.3 (2021), pp. 675–681.

[79] H. Yonezawa, D. Ueda, A. Yamamoto, K. Kageyama, S. L. Walston, T. Nota, K. Murai, S. Ogawa, E. Sohgawa, A. Jogo, D. Kabata, and Y. Miki. "Maskless 2-Dimensional Digital Subtraction Angiography Generation Model for Abdominal Vasculature Using Deep Learning". In: *Journal of Vascular and Interventional Radiology* 33.7 (2022), 845–851.e8.

[80] A. Lindenmayer. "Mathematical Models for Cellular Interactions in Development I. Filaments with One-Sided Inputs". In: *Journal of Theoretical Biology* 18.3 (1968), pp. 280–299.

[81] M. A. Galarreta-Valverde, M. M. G. Macedo, C. Mekkaoui, and M. P. Jackowski. "Three-Dimensional Synthetic Blood Vessel Generation Using Stochastic L-systems". In: *Medical Imaging 2013: Image Processing*. Vol. 8669. SPIE, 2013, pp. 414–419.

[82] J. E. Bresenham. "Algorithm for Computer Control of a Digital Plotter". In: *IBM Systems Journal* 4.1 (1965), pp. 25–30.

[83] G. Melis, C. Dyer, and P. Blunsom. "On the State of the Art of Evaluation in Neural Language Models". In: *International Conference on Learning Representations*. 2018.

[84] K. Musgrave, S. Belongie, and S.-N. Lim. "A Metric Learning Reality Check". In: *Computer Vision – ECCV 2020*. Ed. by A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 681–699.

[85] R. Tibrewala, T. Dutt, A. Tong, L. Ginocchio, R. Lattanzi, M. B. Keerthivasan, S. H. Baete, S. Chopra, Y. W. Lui, D. K. Sodickson, H. Chandarana, and P. M. Johnson. "FastMRI Prostate: A Public, Biparametric MRI Dataset to Advance Machine Learning for Prostate Cancer Imaging". In: *Scientific Data* 11.1 (2024), p. 404.

[86] E. Y. Sidky and X. Pan. "Report on the AAPM Deep-Learning Sparse-View CT Grand Challenge". In: *Medical Physics* 49.8 (2022), pp. 4935–4943.

[87] A. J. Reader, G. Corda, A. Mehranian, C. da Costa-Luis, S. Ellis, and J. A. Schnabel. "Deep Learning for PET Image Reconstruction". In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 5.1 (2021), pp. 1–25.

[88] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. "Focal Loss for Dense Object Detection". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2017, pp. 2999–3007.

[89] K. Yan, X. Wang, L. Lu, L. Zhang, A. P. Harrison, M. Bagheri, and R. M. Summers. "Deep Lesion Graphs in the Wild: Relationship Learning and Organization of Significant Radiology Image Findings in a Diverse Large-Scale Lesion Database". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE, 2018, pp. 9261–9270.

[90] K. Yan, X. Wang, L. Lu, and R. M. Summers. "DeepLesion: Automated Mining of Large-Scale Lesion Annotations and Universal Lesion Detection with Deep Learning". In: *Journal of Medical Imaging* 5.3 (2018), p. 036501.

[91] V. Rawte, S. Chakraborty, A. Pathak, A. Sarkar, S. T. I. Tonmoy, A. Chadha, A. Sheth, and A. Das. "The Troubling Emergence of Hallucination in Large Language Models - An Extensive Definition, Quantification, and Prescriptive Remediations". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino, and K. Bali. Singapore: Association for Computational Linguistics, 2023, pp. 2541–2573.

[92] J. P. Cohen, M. Luck, and S. Honari. "Distribution Matching Losses Can Hallucinate Features in Medical Image Translation". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Ed. by A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger. Cham: Springer International Publishing, 2018, pp. 529–536.

[93] S. Bhadra, V. A. Kelkar, F. J. Brooks, and M. A. Anastasio. "On Hallucinations in Tomographic Image Reconstruction". In: *IEEE Transactions on Medical Imaging* 40.11 (2021), pp. 3249–3260.

[94] M. Genzel, J. Macdonald, and M. März. "Solving Inverse Problems With Deep Neural Networks – Robustness Included?" In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.1 (2023), pp. 1119–1134.

[95] M. Patwari, R. Gutjahr, R. Marcus, Y. Thali, A. F. Calvarons, R. Raupach, and A. Maier. "Reducing the Risk of Hallucinations with Interpretable Deep Learning Models for Low-Dose CT Denoising: Comparative Performance Analysis". In: *Physics in Medicine & Biology* 68.19 (2023), 19LT01.

[96] J. Hsieh, E. Liu, B. Nett, J. Tang, J.-B. Thibault, and S. Sahney. *A New Era of Image Reconstruction: TrueFidelity™*. White Paper. GE Healthcare, 2019.

[97] P. Henzler, V. Rasche, T. Ropinski, and T. Ritschel. "Single-Image Tomography: 3D Volumes from 2D Cranial X-Rays". In: *Computer Graphics Forum* 37.2 (2018), pp. 377–388.

[98] X. Ying, H. Guo, K. Ma, J. Wu, Z. Weng, and Y. Zheng. "X2CT-GAN: Reconstructing CT From Biplanar X-Rays With Generative Adversarial Networks". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, 2019, pp. 10611–10620.

List of Publications

During my four years (May 2021 – May 2025) as a PhD student at the DKFZ, I have contributed to the following publications, three of which were presented in this thesis:

Full papers

[1] **E. Eulig**, A. A. Mastakouri, P. Blöbaum, M. Hardt, and D. Janzing. "Toward Falsifying Causal Graphs Using a Permutation-Based Test". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2025.

[2] **E. Eulig**, F. Jäger, J. Maier, B. Ommer, and M. Kachelrieß. "Reconstructing and Analyzing the Invariances of Low-Dose CT Image Denoising Networks". In: *Medical Physics* 52.1 (2025), pp. 188–200.

[3] **E. Eulig**, B. Ommer, and M. Kachelrieß. "Benchmarking Deep Learning-Based Low-Dose CT Image Denoising Algorithms". In: *Medical Physics* 51.12 (2024), pp. 8776–8788.

[4] L. Duan, **E. Eulig**, M. Knaup, R. Adamus, M. Lell, and M. Kachelrieß. "Training of a Deep Learning Based Digital Subtraction Angiography Method Using Synthetic Data". In: *Medical Physics* 51.7 (2024), pp. 4793–4810.

[5] T. Vöth, T. Koenig, **E. Eulig**, M. Knaup, V. Wiesmann, K. Hörndler, and M. Kachelrieß. "Real-Time 3D Reconstruction of Guidewires and Stents Using Two Update X-ray Projections in a Rotating Imaging Setup". In: *Medical Physics* 50.9 (2023), pp. 5312–5330.

[6] F. Montagna, A. A. Mastakouri, **E. Eulig**, N. Noceti, L. Rosasco, D. Janzing, B. Aragam, and F. Locatello. "Assumption Violations in Causal Discovery and the Robustness of Score Matching". In: *Thirty-Seventh Conference on Neural Information Processing Systems*. 2023.

[7] J. Maier, L. Klein, **E. Eulig**, S. Sawall, and M. Kachelrieß. "Real-Time Estimation of Patient-Specific Dose Distributions for Medical CT Using the Deep Dose Estimation". In: *Medical Physics* 49.4 (2022), pp. 2259–2269.

[8] **E. Eulig**, J. Maier, M. Knaup, N. R. Bennett, K. Hörndler, A. S. Wang, and M. Kachelrieß. "Deep Learning-Based Reconstruction of Interventional Tools and Devices from Four X-ray Projections for Tomographic Interventional Guidance". In: *Medical Physics* 48.10 (2021), pp. 5837–5850.

[9] J. Maier, S. Lebedev, J. Erath, **E. Eulig**, S. Sawall, E. Fournié, K. Stierstorfer, M. Lell, and M. Kachelrieß. "Deep Learning-Based Coronary Artery Motion Estimation and Compensation for Short-Scan Cardiac CT". In: *Medical Physics* 48.7 (2021), pp. 3559–3571.

Conference abstracts & short papers

[10] M. Hammermann, **E. Eulig**, J. Maier, and M. Kachelrieß. "Image-to-Image Translation for Spatial Alignment of Sequential Dual-Energy CT Acquisitions". In: *8th International Conference on Image Formation in X-Ray Computed Tomography*. 2024, pp. 110–113.

[11] A. Kabelac, **E. Eulig**, J. Maier, and M. Kachelriess. "Latent Space Reconstruction and Its Application to CT Detruncation: Latent Detruncation". In: *8th International Conference on Image Formation in X-Ray Computed Tomography*. 2024, pp. 54–57.

[12] G. Kristof, A. Byl, **E. Eulig**, and M. Kachelrieß. "Noise-Augmented Deep Denoising of CT Images". In: *8th International Conference on Image Formation in X-Ray Computed Tomography*. 2024, pp. 86–89.

[13] **E. Eulig**, J. Maier, B. Ommer, and M. Kachelrieß. "May Denoising Remove Structures? How to Reconstruct Invariances of CT Denoising Algorithms". In: *Medical Imaging 2024: Physics of Medical Imaging*. Vol. 12925. SPIE, 2024, pp. 23–28.

[14] **E. Eulig**, B. Ommer, and M. Kachelrieß. "Explainable AI for CT: Analyzing CT Image Denoising Networks by Reconstructing Their Invariances". In: *108th RSNA Scientific Assembly and Annual Meeting*. 2022.

[15] **E. Eulig**, B. Ommer, and M. Kachelrieß. "Reconstructing Invariances of CT Image Denoising Networks Using Invertible Neural Networks". In: *7th International Conference on Image Formation in X-Ray Computed Tomography*. Vol. 12304. SPIE, 2022, pp. 169–173.

[16] J. Maier, L. Jordan, **E. Eulig**, F. Jäger, S. Sawall, M. Knaup, and M. Kachelrieß. "Learning CT Scatter Estimation without Labeled Data: A Feasibility Study". In: *7th International Conference on Image Formation in X-Ray Computed Tomography*. Vol. 12304. SPIE, 2022, pp. 675–679.

[17] T. Vöth, T. König, **E. Eulig**, M. Knaup, V. Wiesmann, K. Hörndler, and M. Kachelrieß. "3D Reconstruction of Stents and Guidewires in an Anthropomorphic Phantom from Three X-Ray Projections". In: *7th International Conference on Image Formation in X-Ray Computed Tomography*. Vol. 12304. SPIE, 2022, pp. 223–229.

[18] T. Vöth, T. König, **E. Eulig**, M. Knaup, K. Hörndler, and M. Kachelrieß. "Real-Time 3D Reconstruction of Multiple Guidewires at Dose Values of Conventional 2D Fluoroscopy". In: *Medical Imaging 2022: Physics of Medical Imaging*. Vol. 12031. SPIE, 2022, pp. 405–409.

Appendix

A Supplementary Material: Benchmarking Deep Learning-Based Low-Dose CT Image Denoising Algorithms

The following pages contain the supplementary material of our paper on benchmarking deep learning-based low-dose CT image denoising algorithms [63].

Appendix: Benchmarking Deep Learning-Based Low-Dose CT Image Denoising Algorithms

A. Implementation and verification of algorithms

The eight algorithms implemented in this study were all implemented in PyTorch and will be made publicly available together with our benchmark suite. In Fig. A.1 we provide a flowchart of how we implemented and verified a given denoising algorithm.

Implementation For implementation (Fig. A.1; top) of an algorithm we first checked if an implementation was open-sourced by the authors of the original paper, which was the case for the four newest algorithms considered in this study (Tab. A.1). For two of these, implementations were in TensorFlow and had to be translated to PyTorch. For two other algorithms, we found open-source implementations by a third party. Only for two (relatively simple and old) algorithms, namely CNN-10 and ResNet, no implementation was found, and we had to solely rely on the information provided in the original paper for our implementation. For all other algorithms we closely compared results and implementation of individual modules in architecture (*e.g.*, residual blocks) and training pipeline (*e.g.*, loss computation).

Verification For all algorithms we performed unit testing of each module. We then checked whether the authors performed experiments on the *LDCT Image and Projection dataset* or a subset thereof (*e.g.*, *NIHAAPM-Mayo Clinic LDCT Grand Challenge*). If this was the case, we verified that our results (with similar hyperparameters as described by the authors) were quantitatively similar to those reported in the paper. If this was not the case, we had to rely on a qualitative evaluation to verify that our implementation was correct.

B. Usage of the benchmark suite

We provide our benchmark suite as a Python package that can be easily used by researchers to evaluate their algorithms. The package contains the following components:

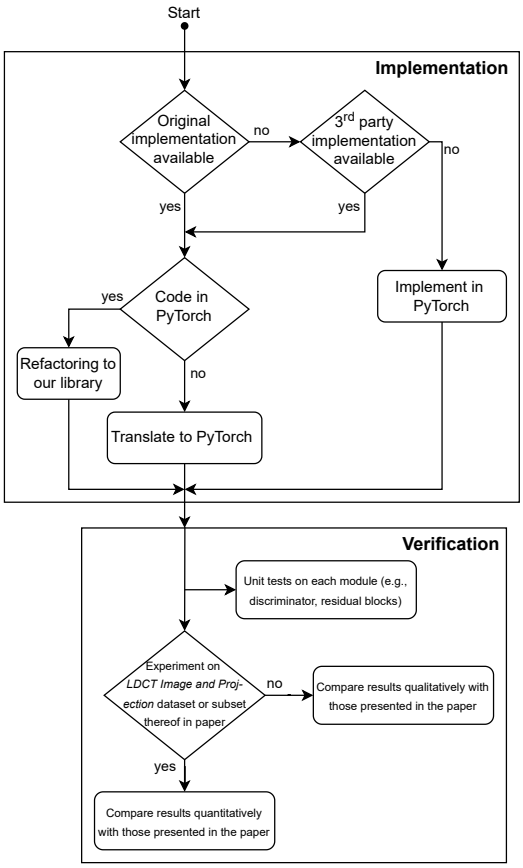


Figure A.1: Flowchart of the implementation and verification process performed for each algorithm considered in this study.

- **Data loader** for the training, validation, and test sets we used of the *LDCT Image and Projection dataset*. This includes preprocessing steps such as normalization and random cropping and makes sure that each algorithm is trained and evaluated using the same data.
- **Base trainer** that provides a base class that provides methods for training, validating and logging of a given algorithm. Researchers can inherit from this class to easily implement their own algorithms in our benchmark suite.
- **Hyperparameter optimization pipeline** which can be used to optimize hyperparameters for a given algorithm using Bayesian optimization. Researchers can provide command line arguments to their algorithm through which hyperparameters are defined. After specifying hyperparameters to be optimized and their priors in a YAML file the pipeline will optimize these arguments for the given algorithm. An example of such a configuration file is shown in Tab. B.2.

Table B.2: Example of a hyperparameter optimization configuration file for CNN-10.

```
method: bayes
metric:
  goal: maximize
  name: SSIM
name: hpopt-cnn10
parameters:
  adam_b1:
    value: 0.9
  adam_b2:
    value: 0.999
  data_norm:
    value: meanstd
  data_subset:
    value: 1.0
  datafolder:
    value: /path/to/LDCTData
  iterations_before_val:
    value: 1000
  lr:
    distribution: log_uniform_values
    max: 0.01
    min: 1.0e-05
  max_iterations:
    distribution: int_uniform
    max: 100000
    min: 1000
  mbs:
    distribution: int_uniform
    max: 128
    min: 2
  num_workers:
    value: 4
  optimizer:
    value: adam
  patchsize:
    distribution: int_uniform
    max: 128
    min: 32
  seed:
    value: 1332
  trainer:
    value: cnn10
```

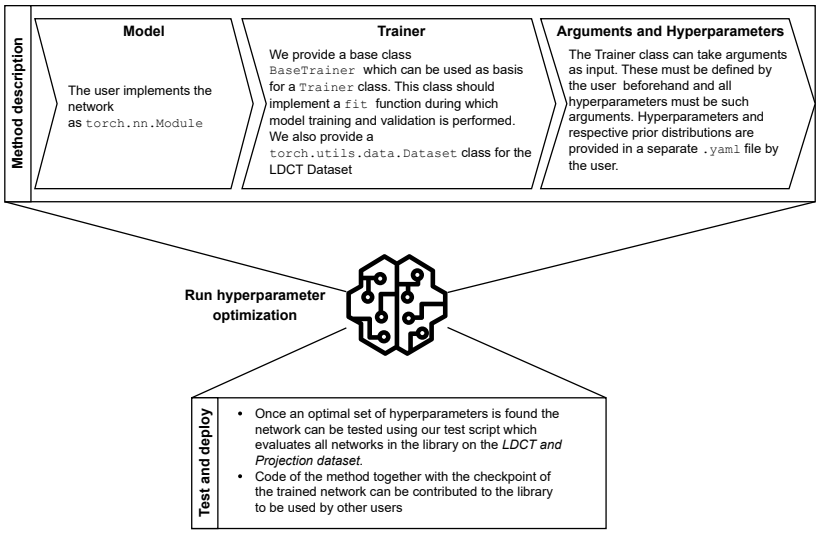


Figure B.2: Flowchart of our benchmark pipeline.

- **Evaluation scripts** to evaluate a given algorithm using the same metrics and settings as in our study. This includes the evaluation of radiomic feature similarity and lesion annotations.

- **Model hub** in which pretrained models of all algorithms considered in this study are provided and to which researchers can contribute their own models.

A flowchart of our benchmark pipeline is shown in Fig. B.2. We provide more information in the documentation of the package.

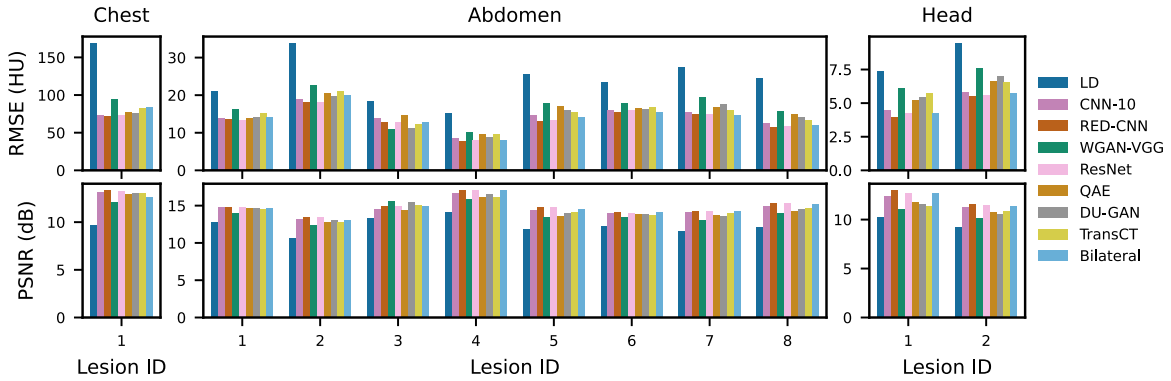


Figure C.3: Quantitative evaluation for each of the lesions in the test set. Lesion IDs correspond to # provided in Figs. C.4 to C.6.

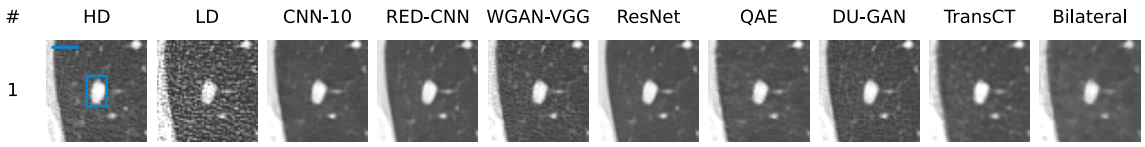


Figure C.4: Solid, non-calcified lung nodule on chest exam from the test set. Blue line indicates one centimeter, blue bounding-box indicates the lesion annotation.

C. Evaluation on lesions

We show reconstruction results on all lesions and algorithms in Figs. C.4 to C.6 and present quantitative evaluation of all individual lesions in Fig. C.3. The quantitative evaluation is performed inside the bounding box annotation shown in blue in Figs. C.4 to C.6.

D. Line profile analysis

For the chest exam (Fig. D.7a) and abdomen exam (Fig. D.7b), we observe that all algorithms reduce the noise compared to the LD reconstruction (blue curve; top-left line plot) and stay closer to the high dose reconstruction (black curve in all plots). However, some algorithms fail to recover sharp edges in the line profile (CNN-10, QAE, TransCT, and Bilateral for the chest exam and Bilateral for the abdomen exam). We find that RED-CNN, ResNet, WGAN-VGG, and DU-GAN perform best in this regard for both exams. For the head exam (Fig. D.7c), while all algorithms reduce noise compared to the LD reconstruction, they all fail to recover the sharp edges in the line profile except TransCT.

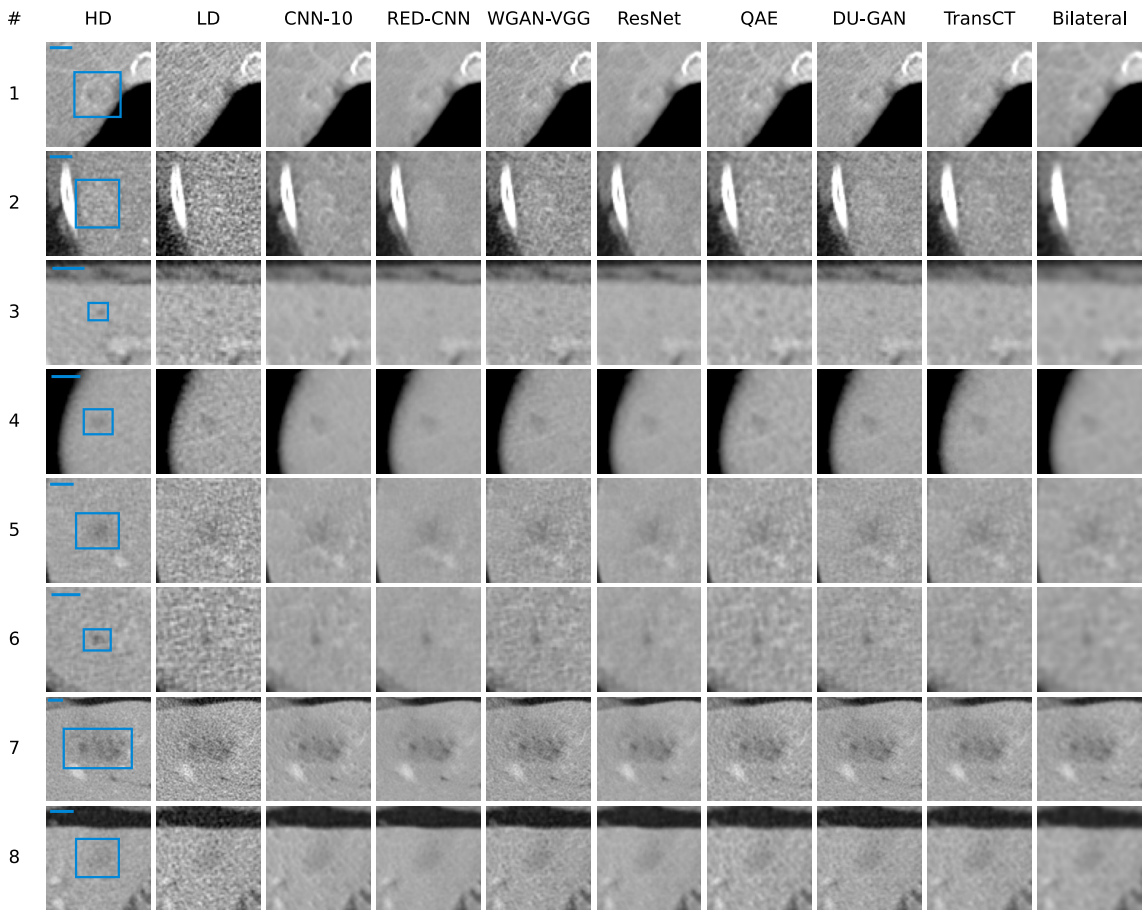


Figure C.5: Pathologies on abdomen exams from the test set. Blue line indicates one centimeter, blue bounding-box indicates the lesion annotation. (1, 2, 5, 7, 8): metastasis, (3, 6): cyst, 4: hemangioma.

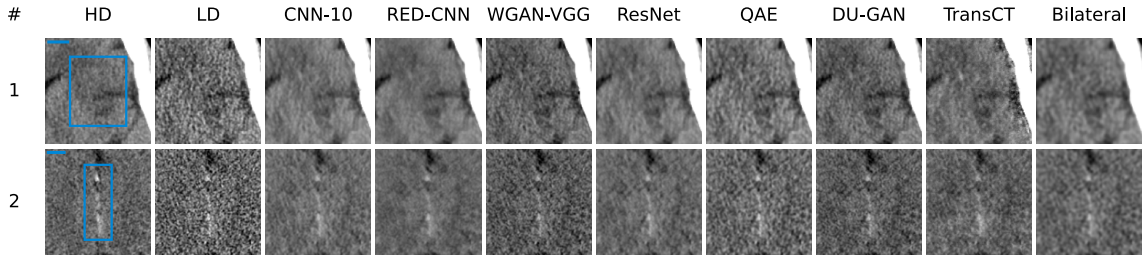
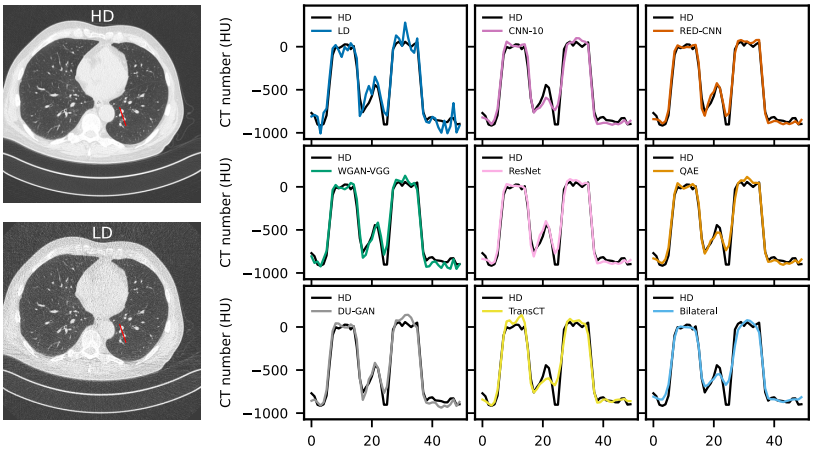
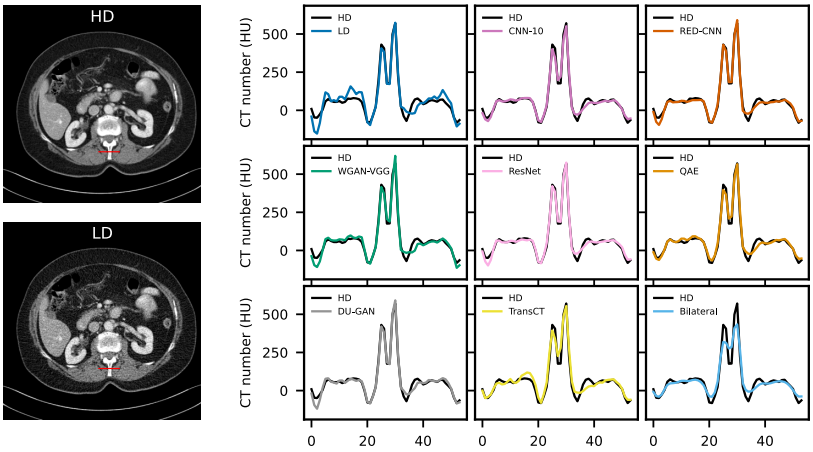


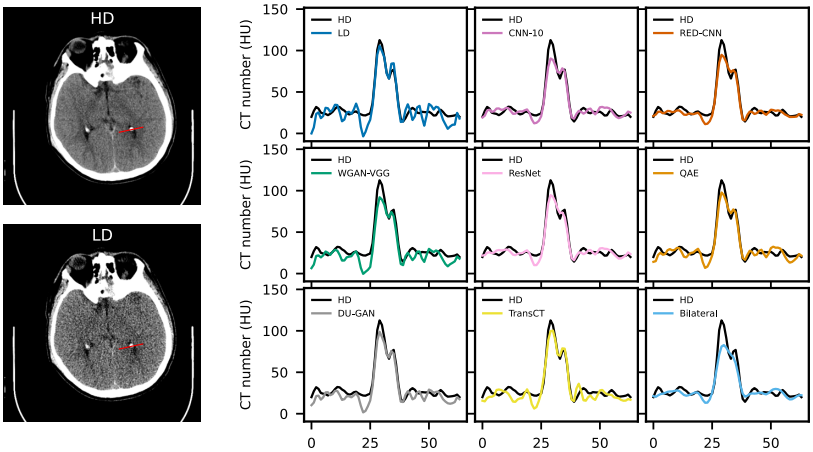
Figure C.6: Pathologies on head exams from the test set. Blue line indicates one centimeter, blue bounding-box indicates the lesion annotation. 1: acute frontal infarct, 2: traumatic subarachnoid hemorrhage.



(a) Line profile for chest exam



(b) Line profile for abdomen exam



(c) Line profile for head exam

Figure D.7: Line profiles for chest (a), abdomen (b), and head (c) exams. The line along which the profile is computed is indicated in red in the high dose and low dose image.

B Supplementary Material: Reconstructing and Analyzing the Invariances of Low-Dose CT Image Denoising Networks

The following pages contain the supplementary material of our paper on invariances of low-dose CT image denoising networks [69].

Appendix: Reconstructing and Analyzing the Invariances of Low-Dose CT Image Denoising Networks

A. Training details

A.1. Denoising networks

For all denoising methods, we determine the optimal hyperparameters via Bayesian optimization on the *Low Dose CT Image and Projection Dataset*[31]. For more details we refer the reader to Eulig et al., 2024[32] and list the hyperparameters here for reproducibility: **CNN-10** was trained on patches of size 92×92 px with a mini-batch size of 68, and using the Adam optimizer with learning rate 1.6×10^{-4}. **RED-CNN** was trained on patches of size 128×128 px, with a mini-batch size of 73, and using the Adam optimizer with learning rate 9.6×10^{-5}. **WGAN-VGG** was trained on patches of size 77×77 px, with a mini-batch size of 77, and using the Adam optimizer with learning rate 7.1×10^{-5} and $\beta_1 = 0.33$. The perceptual loss is evaluated after the last convolutional layer of a pretrained VGG-19 network and weighted with $\lambda_{\text{perc}} = 0.689$. **DU-GAN** was trained on patches of size 128×128 px, with a mini-batch size of 92, and using the Adam optimizer with learning rate 1.2×10^{-5} and $\beta_1 = 0.65$. As weighting parameters for the generator loss we used $\lambda_{\text{img}} = 1.0, \lambda_{\text{grd}} = 27.8, \lambda_{\text{adv}} = 0.1$. We use the same CutMix regularization as described in the original publication.

A.2. Conditional variational autoencoder

Our conditional autoencoder uses an ImageNet-pretrained ResNet-50[38] as encoder that predicts $\mu^{(i)}, \sigma^{(i)} \in \mathbb{R}^M$ with $M = 256$. Latent variables are mapped to image space using a decoder based on BigGAN[39]. To improve reconstruction quality we use a mixture of pixelwise, perceptual, and adversarial loss for the reconstruction term $\mathcal{L}_{\text{rec}}$ in Eq. (4):

$$\begin{aligned}\mathcal{L}_{\text{rec}} &= \mathcal{L}_{\text{adv}} + \lambda_{\text{pix}}\mathcal{L}_{\text{pix}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} \\ &= \mathbb{E}_i \left[-C\left(\hat{x}^{(i)} | \hat{y}^{(i)}\right) + \lambda_{\text{pix}} |x^{(i)} - \hat{x}^{(i)}| + \lambda_{\text{perc}} \frac{1}{L} \sum_{l=1}^{L} |\phi_l^{\text{VGG}}(x^{(i)}) - \phi_l^{\text{VGG}}(\hat{x}^{(i)})| \right], \end{aligned} \tag{11}$$

with $\lambda_{\text{pix}} = 1.2 \times 10^{-3}, \lambda_{\text{perc}} = 0.12$ determined through a randomized search. The perceptual loss is evaluated before each of the first $L = 4$ max-pooling operations of an

ImageNet-pretrained VGG-16 ϕ^{VGG}. The adversarial loss is evaluated using a critic with six blocks where each block consists of Conv2D($f_{\mathrm{in}}, f_{\mathrm{out}}, \mathrm{kernel\_size} = 4, \mathrm{stride} = 2$) $\rightarrow$ InstanceNorm $\rightarrow$ LeakyReLU(negative_slope $= 0.2$) with $f_{\mathrm{in}} = 2, 64, 128, 256, 512, 512$ and $f_{\mathrm{out}} = 64, 128, 256, 512, 512, 512$. Lastly, a single 1×1 convolution followed by a fully-connected layer Linear$(16, 1)$ is used to map to a scalar output.

Both generator and critic are trained with a batch size of 64 using an Adam optimizer with $\beta_1 = 0.5$ and a learning rate of 1×10^{-5} for the generator and 4×10^{-5} for the critic.

A.3. Conditional invertible neural network

The cINN t_ξ consists of 12 invertible blocks, each of which consists of an affine coupling layer[41], ActNorm[43], and a shuffling layer. The output of the denoising network is first embedded using a shallow CNN and then concatenated to each of the fully connected networks in the coupling layers to realize the conditioning. Since our architectural design closely follows the design of the cINN in Rombach et al., 2021[14] we refer the reader to the original publication for further details.

We train the cINN with the loss function in Eq. (6) using the Adam optimizer and a learning rate of 1×10^{-5} and a mini-batch size of 128.

A.4. Learned embedding

As DML model h_ν we use the feature encoder of an ImageNet-pretrained ResNet-50 with a single fully-connected layer after the average pooling, mapping to the 32-dimensional embedding space. All embeddings are ℓ_2 normalized. During training, we sample triplets $(a, p, n) := (x^{(i)}, \tilde{x}^{(i,k)}, \tilde{x}^{(j,l)})$ using the online version of the semi-hard triplet mining strategy[45], *i.e.*, for all positives and negatives in a batch we only use triplets for which the negative is further away from the anchor than the positive but violates the margin α:

$$d(h(a), h(p)) < d(h(a), h(n)) < d(h(a), h(p)) + \alpha\,. \tag{12}$$

As most related works, we use the squared euclidian distance $d(x, y) = \|x - y\|_2^2$ as distance metric during training. Note that for ℓ_2-normalized embeddings, using the cosine similarity Eq. (10) for analyzing sampled invariances does not alter the ranking of similar/dissimilar

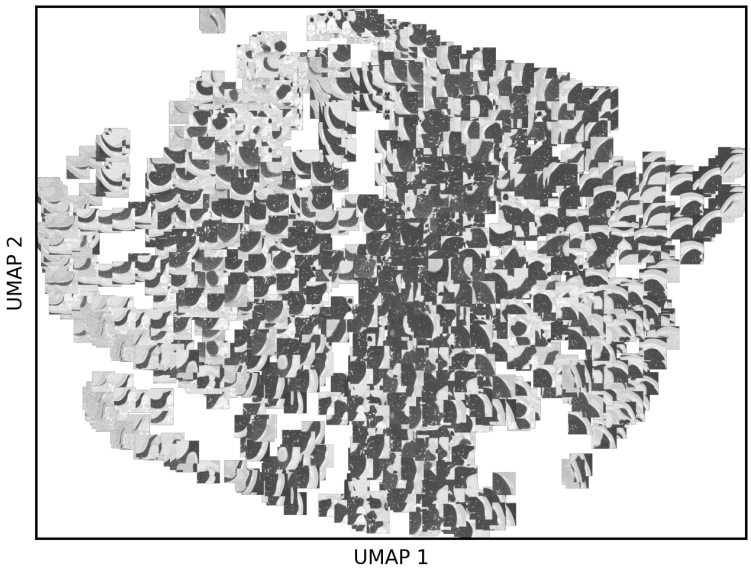


Figure B.1: UMAP visualization of the 256-dimensional VAE latent space for 1,000 samples $x^{(i)}$ from the test dataset, encoded using the unconditional VAE.

samples shown in Fig. 7 since

$$\|x - y\|_2^2 = 2 - 2\cos(x, y) \quad \forall \|x\|_2 = \|y\|_2 = 1\,. \tag{13}$$

We train the model using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 128. The margin α is set to 0.2.

B. Additional results

VAE latent space To further test our hypothesis that the VAE latent space is dominated by anatomical information, we visualize its latent space in Fig. B.1. This is done by sampling 1,000 images from the test dataset, predicting the mean using the VAE encoder and applying UMAP[53] to reduce the 256 latent dimensions down to two. Here we find that similar anatomical structures and regions are, independently of their noise structure and realization, embedded close together in the latent space.

Distributions of $S_{\mathbf{VAE}}$ and $S_{\mathbf{DML}}$ We show violinplots of the cosine similarities S_{VAE} and S_{DML} for all 1000 random samples from the test set in Fig. B.2.

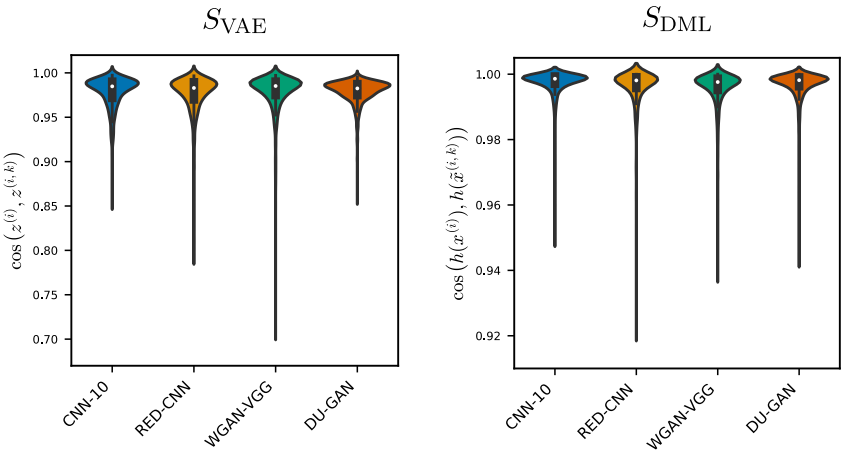


Figure B.2: Violinplots of the cosine similarities S_{VAE} and S_{DML} for all 1000 random samples from the test set.

Conditional VAE reconstructions The ability of the conditional VAE to faithfully reconstruct the input is to some extent dependent on the conditioning $\hat{y} = f_\theta(x)$ and therefore on the denoising network f_θ. We show reconstruction capabilities for all four conditional VAEs (conditioned on the four denoising networks, respectively) as well as for an unconditonal VAE in Fig. B.3. We find that while there are, as expected, small differences in the reconstructions, all conditional VAEs are able to reconstruct the low dose images very well and are able to capture all anatomical structures. In particular, all conditional VAEs perform much better than the unconditional VAE used in previous work[30].

Case study: Algorithm with strong invariances by design Since all four networks investigated in our study showed relatively little invariances to anatomical structures, we conduct a case study where we reconstruct the invariances of a dummy algorithm that has strong localized invariances by design. Let $M \in \{0,1\}^{n\times m}$ be a binary mask of same size as the input images $x^{(i)} \in \mathbb{R}^{n\times m}$ then the output image $\hat{y}^{(i)} \in \mathbb{R}^{n\times m}$ of this algorithm is given as

$$\hat{y}^{(i)} = x^{(i)} \odot (1 - M)\,, \tag{14}$$

with $\odot$ denoting the element-wise product, *i.e.* all pixels in $x^{(i)}$ for which M is one are set to zero. In our experiments we use a mask M that is a square of size 30×30 pixels centered in the image and denote the resulting algorithm as **Center-Inv**. We expect this algorithm to be, by design, invariant w.r.t. input information within this center square but not invariant to

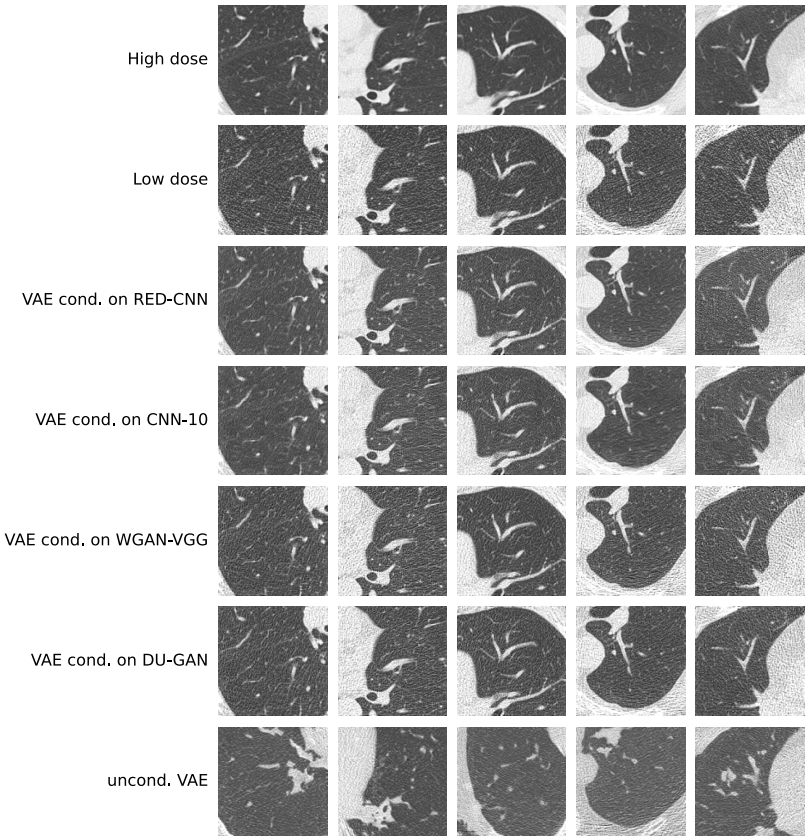


Figure B.3: High dose, low dose, and VAE reconstructions for all four conditional VAEs used in this work (conditioned on the four denoising networks, respectively). Additionally, we show the reconstructions of an unconditional VAE (as used in previous work[30]) for comparison. $C = -600\,\text{HU}$, $W = 1500\,\text{HU}$.

anything outside the square. For this algorithm we train an additional conditional VAE and cINN as described in App. A.2. and A.3.. We show one random example of reconstructed invariances for Center-Inv in Fig. B.4. We find that our method is able to successfully reconstruct the invariances within the center square while showing close-to-zero invariances outside the square. Within the center square, differences between sampled invariances are larger than they are for the other four algorithms, as expected (Fig. B.4; rightmost column). We also perform a quantitative evaluation using the mean absolute difference (MD), mean cosine similarity in the VAE latent space (S_{VAE}), and mean cosine similarity in the learned embedding space (S_{DML}), similar to the evaluation in Sec. III.D. and show the results in Tab. B.1. We find that Center-Inv has significantly more anatomical invariances compared to the other algorithms as measured by the S_{VAE} and S_{DML}. Furthermore, we find that the mean absolute difference (MD) between reconstructed invariances and input images is

Table B.1: Quantitative evaluation of invariances using the mean absolute difference (MD), mean cosine similarity in the VAE latent space (S_{VAE}), and mean cosine similarity in the learned embedding space (S_{DML}). The first four rows are identical to Tab. 2 and are shown for comparison.

Invariances	MD ↑ noise + content	S_{VAE} ↓ content	S_{DML} ↓ content
CNN-10	182 ± 67	0.978 ± 0.020	0.997 ± 0.004
RED-CNN	$\mathbf{191 \pm 67}$	0.976 ± 0.022	0.996 ± 0.007
WGAN-VGG	158 ± 64	0.979 ± 0.021	0.996 ± 0.006
DU-GAN	178 ± 70	0.979 ± 0.013	0.997 ± 0.004
Center-Inv	46 ± 17	$\mathbf{0.960 \pm 0.032}$	$\mathbf{0.983 \pm 0.041}$

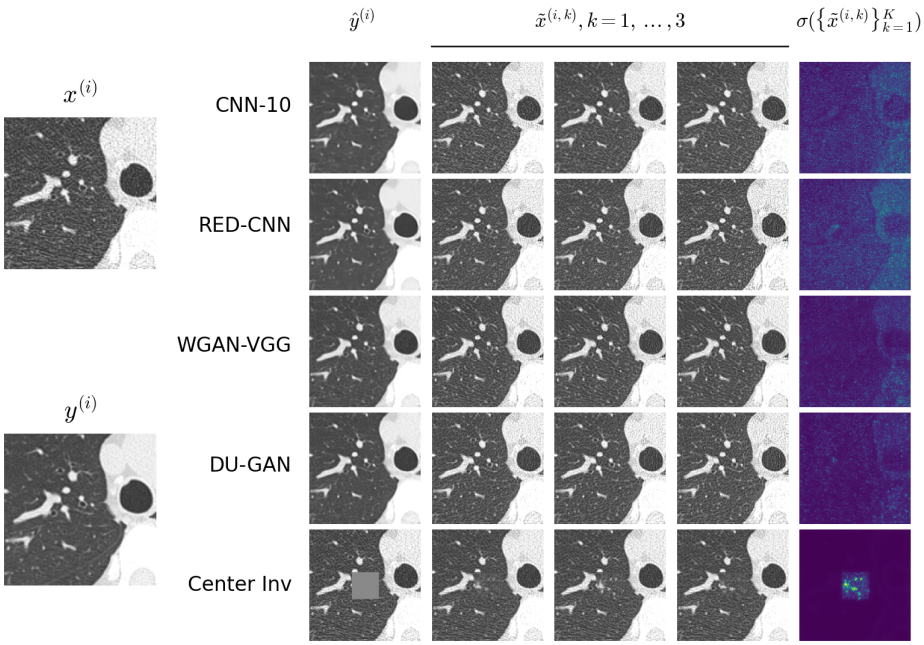


Figure B.4: Three reconstructed invariances for a random sample from the test set. For each of the methods we also show the denoised image $\hat{y}$ and the standard deviation map over all $K = 100$ sampled invariances. For CT Images: $C = -600\,\text{HU}$, $W = 1500\,\text{HU}$, for standard deviations: $C = 0\,\text{HU}$, $W = 300\,\text{HU}$.

lower for Center-Inv than for the other algorithms. This can be explained by our method successfully predicting almost zero invariances for all pixels outside the mask which account for $\approx 94.5\%$ of all pixels in the image.

Reconstructed invariances We present additional invariances reconstructed for six random samples from the test set in Figs. B.5 to B.10.

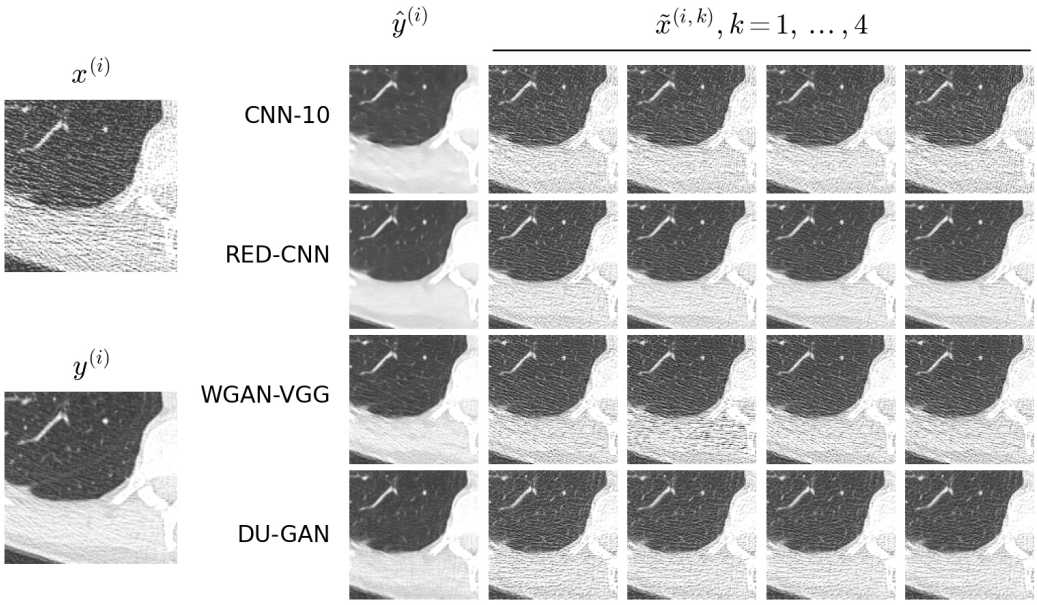


Figure B.5: Four reconstructed invariances for a random sample from the test set. For each of the methods we also show the denoised image $\hat{y}$. $C = -600\,\text{HU}$, $W = 1500\,\text{HU}$.

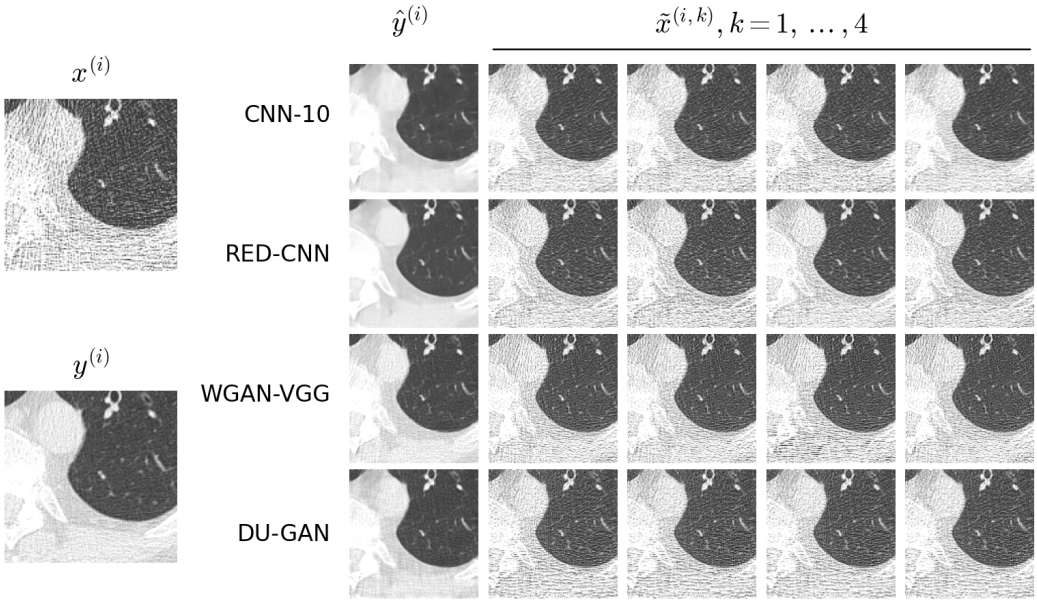


Figure B.6: Four reconstructed invariances for a random sample from the test set. For each of the methods we also show the denoised image $\hat{y}$. $C = -600\,\text{HU}$, $W = 1500\,\text{HU}$.

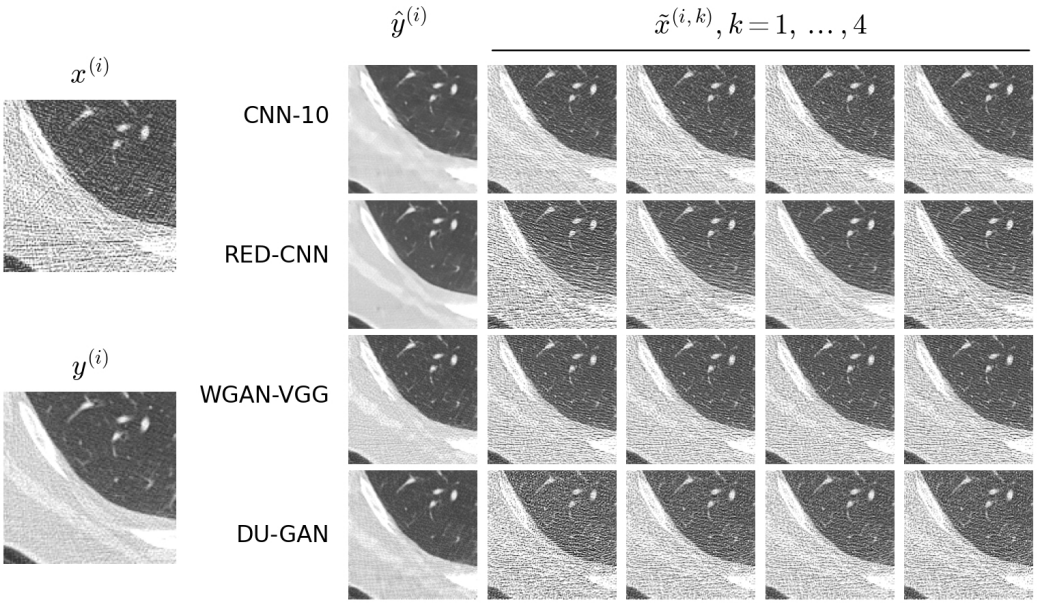


Figure B.7: Four reconstructed invariances for a random sample from the test set. For each of the methods we also show the denoised image $\hat{y}$. $C = -600\,\text{HU}$, $W = 1500\,\text{HU}$.

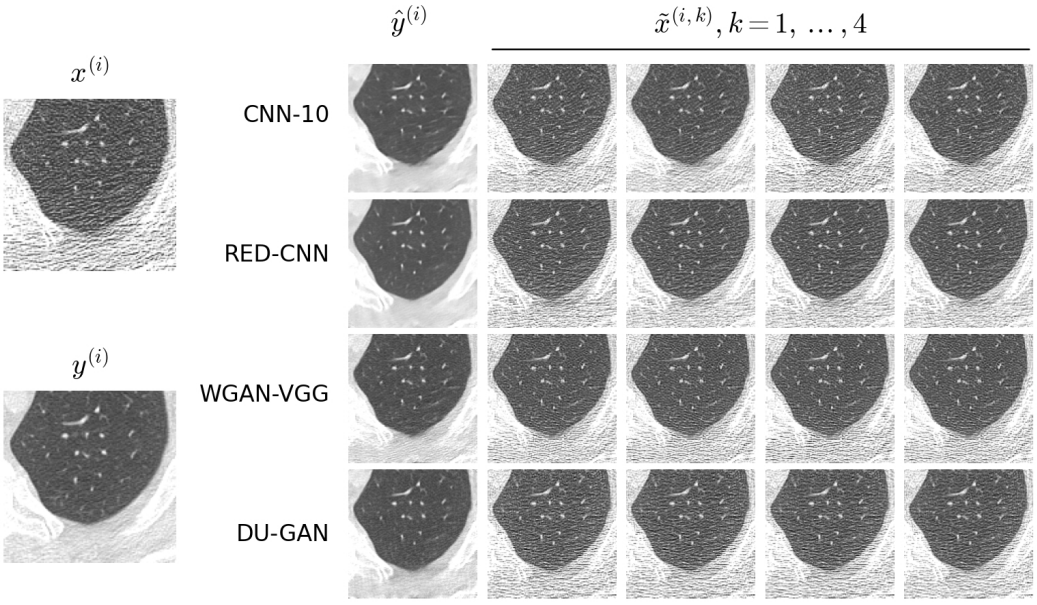


Figure B.8: Four reconstructed invariances for a random sample from the test set. For each of the methods we also show the denoised image $\hat{y}$. $C = -600\,\text{HU}$, $W = 1500\,\text{HU}$.

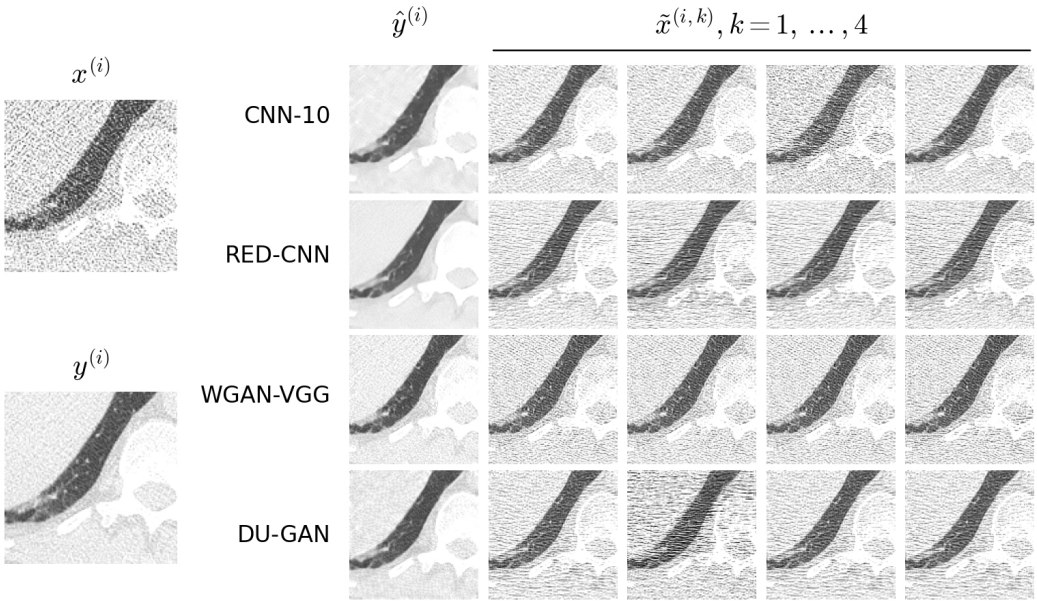


Figure B.9: Four reconstructed invariances for a random sample from the test set. For each of the methods we also show the denoised image $\hat{y}$. $C = -600\,\text{HU}$, $W = 1500\,\text{HU}$.

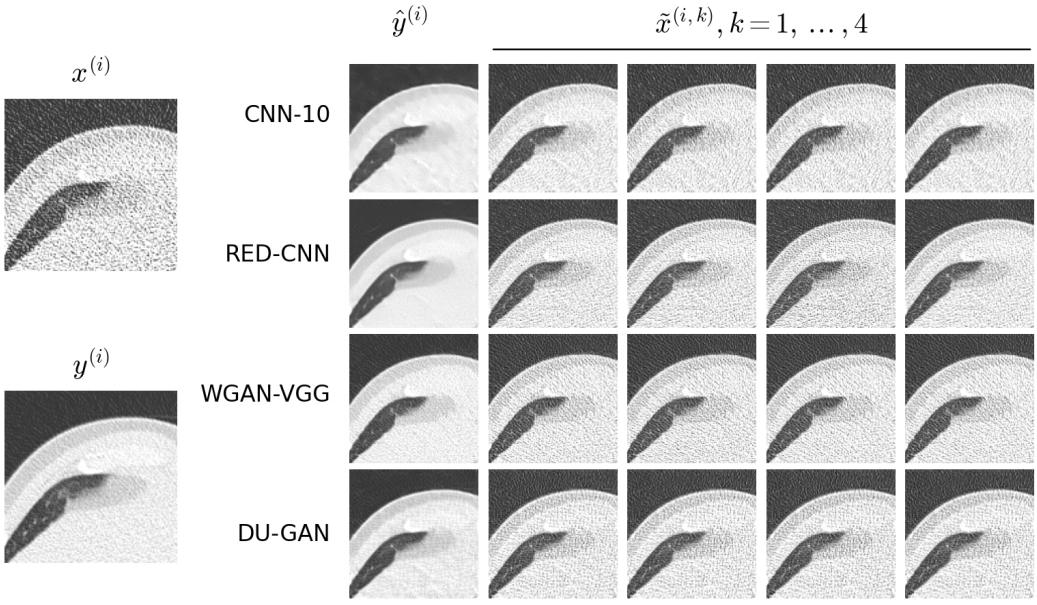


Figure B.10: Four reconstructed invariances for a random sample from the test set. For each of the methods we also show the denoised image $\hat{y}$. $C = -600\,\text{HU}$, $W = 1500\,\text{HU}$.

C Supplementary Material: Training of a Deep Learning-Based Digital Subtraction Angiography Method Using Synthetic Data

The following pages contain the supplementary material of our paper on synthetic data for the training of deep learning-based DSA [72].

APPENDIX A: NETWORK AND TRAINING

More detailed architecture of networks and parameter settings are presented in this section.

A.1 | Networks

The U-net architecture is shown in Figure A1. It contains five max-pooling layers and five transposed convolutional layers, all preceded by two standard convolutional layers. The last transposed convolutional layer is followed by two standard convolutional layers and a final convolutional layer. The kernel size is 4×4 in all transposed convolutional layers and 3×3 in all standard convolutional layers. The leakyReLU activation function with a slope of 0.01 is applied after each convolutional layer. The numbers of filters for the standard convolutional layers are 64, 128, 256, 512, 1024, 1024, 512, 256, 128, 64, and 64 respectively. The output of the network is obtained after the final convolutional layer with a convolutional kernel size of 1×1 and a number of filters of 1.

The generator network of the GAN adopts a U-net architecture similar to the one shown in Figure A1, but with different numbers of filters for the standard convolutional layers. Specifically, the number of filters for the standard convolutional layers is set as 16, 32, 64, 128, 256, 256, 128, 64, 32, 16, and 16, respectively. The discriminator network is shown in Figure A2, which contains seven basic convolutional feature extraction layers and a fully connected layer. Each convolutional layer is followed by instance normalization[24] and leakyReLU activation with a slope of 0.2 except for the last convolutional layer. The kernel size for the first five convolutional layers is 4×4 and the stride is 2. The kernel size for the sixth convolutional layer is 3×3 and the stride is 1. The last convolutional layer has a convolutional kernel size of 1×1 and a stride of 1. The number of filters are 32, 64, 128, 256, 512, 512, 1, and 1. The output of the fully connected layer is a score indicating the likelihood that the input images include a DSA image rather than a DDSA image generated by the generator network. The overview of the conditional GAN procedure is demonstrated in Figure A3.

A.2 | Implementation details

Unless stated otherwise, networks f were trained with a MAE loss

$$L_{\text{pix}}(f) = \mathbb{E}_x ||f(x) - y||_1 \,, \tag{A1}$$

where x are sampled x-ray images and y and $f(x)$ are the corresponding ground truth DSA images and network outputs, respectively.

For networks trained in a GAN setting, we used the following loss[21] the discriminator D:

$$\begin{aligned} L_D &= \mathbb{E}_x[D(x, G(x))] - \mathbb{E}_x[D(x, y)] \\ &\quad + \lambda \mathbb{E}_{\hat{x}}[(||\nabla_{\hat{x},y} D(\hat{x}, y)||_2 - 1)^2] \,, \end{aligned} \tag{A2}$$

where $G(x)$ are outputs of the generator G, the points of $\hat{x}$ are sampled uniformly along straight lines between pair points sampled from x and $G(x)$, and λ is set to 10.

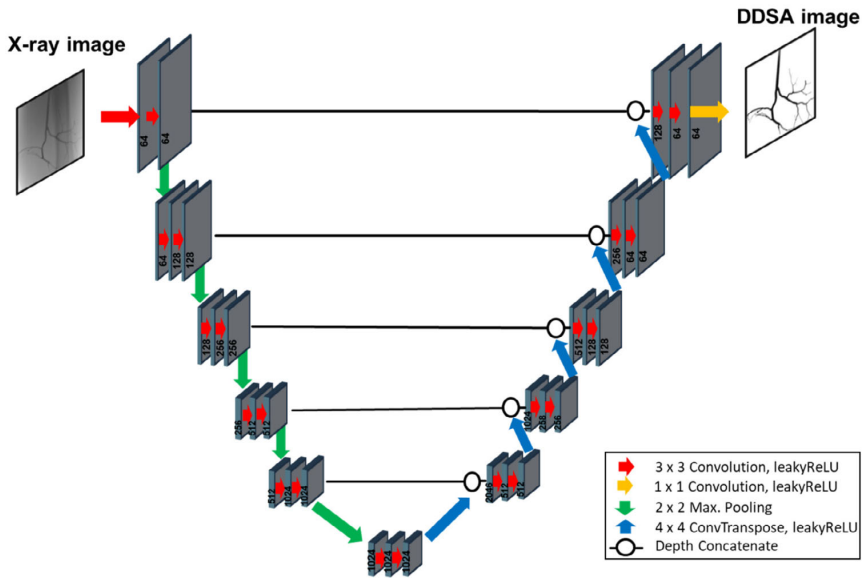


FIGURE A1 Network architecture of U-net.

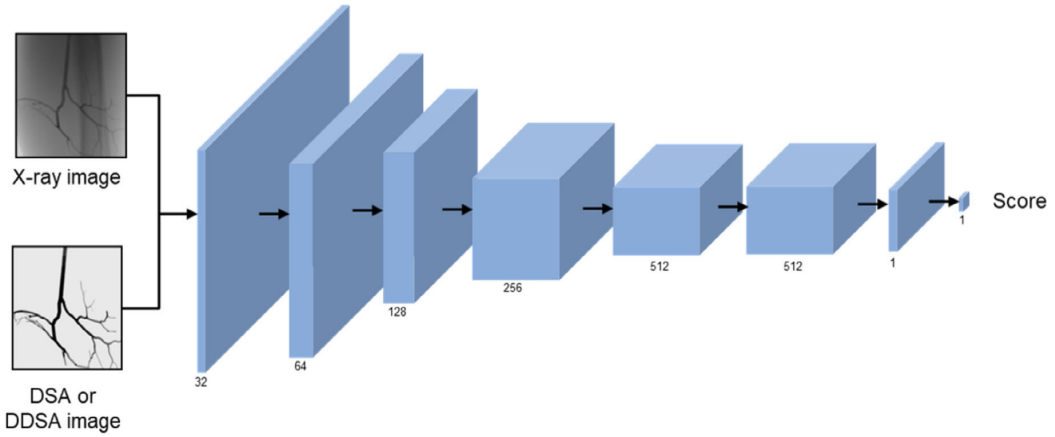


FIGURE A2 The discriminator model of conditional GAN.

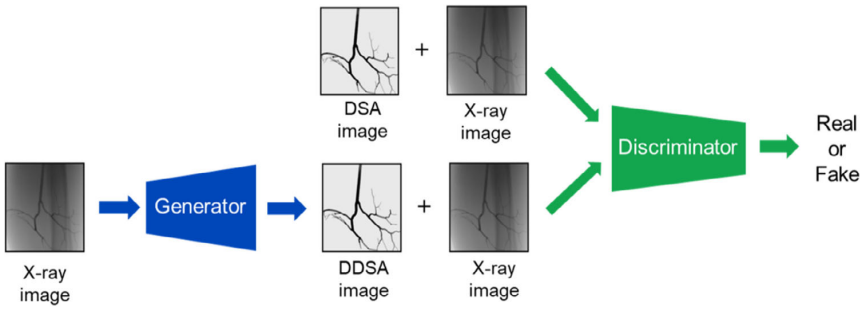


FIGURE A3 Procedure of conditional GAN.

The loss for the generator G is composed of a standard adversarial loss and the pixel-wise loss defined above:

$$L_G = L_{\text{pix}}(G) - \beta \mathbb{E}_x[D(x, G(x))] \,, \tag{A3}$$

where β is set to 10.

Training of the neural network was performed using randomly-cropped image patches with a size of 384 × 384 pixels. The training was performed for 75 epochs. Every epoch contained 36 800 randomly-picked training image patches and 4800 randomly-picked test image patches. To increase the networks' ability to generalize, data augmentation was employed to create more training data online. The data augmentation consisted of (1) horizontal and vertical flipping both with a probability of 0.5, (2) rotating around the image center with a maximum deviation of $\pm 180°$, (3) shearing around the image center with a maximum deviation of $\pm 20°$, (4) scaling image size by a factor between 0.5 and 1.5, (5) average blur with kernel sizes randomly between 0 and 5 and a probability of 0.8, (6) piecewise affine transformations with a range from 0 to 0.05 and a probability of 0.8, and (7) scaling of image intensity randomly between factor 0.7 to 1.3 and a probability of 0.8.

All networks were trained with Adam[25] optimizer using an initial learning rate of 0.0001, momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$. For training in the GAN setting we used momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$. To ensure the convergence of the adversarial learning, the generator was trained once after optimizing the discriminator four times. The training batch size was set to 32 for both networks.

APPENDIX B: ADDITIONAL EXPERIMENTS

B.1 | Effect of mimicking the flow of the contrast agent

Figure B1 and Table B1 show the visual and evaluation results on clinical datasets with and without mimicking contrast flow to synthetic DSA images. The results show that mimicking contrast flow can facilitate network to predict better DSA images.

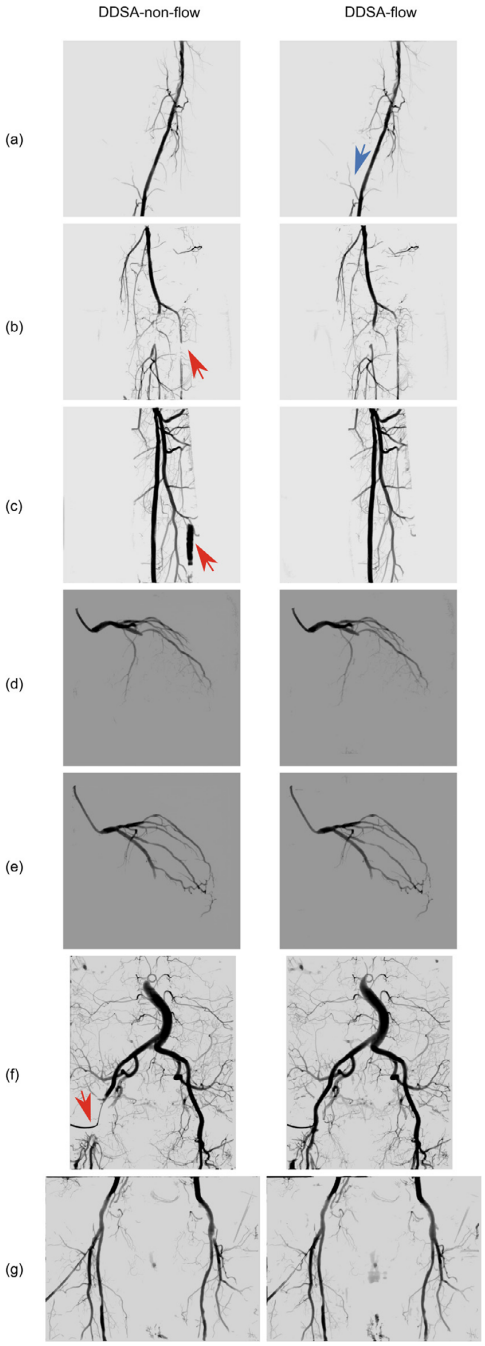


FIGURE B1 Visual results for analyzing the effects of contrast flow simulation: (a)–(g) represent different test data. Left to right: results of U-net trained without mimicking contrast flow (DDSA-non-flow), and results of U-net trained with mimicking contrast flow (DDSA-flow). Blue arrows indicate more vessels extracted by DDSA models. Red arrows indicate missing vessels and wrongly inferred vessels by DDSA models. DDSA, deep DSA.

B.2 | Effect of the training set size of CT data

We randomly sampled different percentage of CT data from the original dataset containing 105 CT series to create training sets of varying sizes (25%, 50%, 100% of the original dataset). Finally, three CT datasets with 25, 52, and 105 CT data were obtained, respectively. For a fair comparison with clinical data where each datum contains only one mask image, only one of the 72 projection images of each CT data was randomly selected as the mask image. All synthetic vascular projection images were used to generate synthetic DSA images. We trained the U-net model with pixel-loss and evaluated their performance on the three clinical test datasets.

The visual and quantitative results are shown in Figure B2 and Table B2, respectively. All three training sets that use synthetic data produce clearer DSA images and predict more complete vessels on abdominal and cardiac data compared to the training set that uses clinical data. Training on the training set containing 25 CT data produces more false positive vessels on the leg data compared to the other two training sets. Compared to the training set containing 62 clinical data, the training set containing 52 CT data lost more vessels on the difficult test data (Figure B2b),but has higher performance on the other two data. The above results further demonstrate the superiority of our synthetic data for training the DDSA model. The poor performance on the specific test data can be attributed to the independent selection of both the mask images and synthetic vascular projection images during training, as discussed in Section 4.

B.3 | Effect of anatomical sites

The training CT dataset of the synthetic dataset includes different anatomical sites, while the clinical training dataset only consists of data from legs. To test the generalizability of our methods, we created a new dataset by compiling 17 projected CT series of legs and used them as mask images to train the U-net model. The test results, shown in Figure B3, demonstrate that the model trained with this new dataset predicts more detailed vessels on the abdomen and cardiac datasets, and shows comparable results on the leg dataset when compared to the model trained on the clinical DSA dataset. This model produces more artifacts when testing on the abdomen and cardiac datasets due to the limited mask training data compared to the one trained with all CT data as mask images. The quantitative results, as depicted in Table B3, demonstrate that the new results outperform those obtained by training on clinical data and, in some instances, even those obtained by training on all CT data. The above results indicate that the network trained on our synthetic dataset has better generalization and robustness.

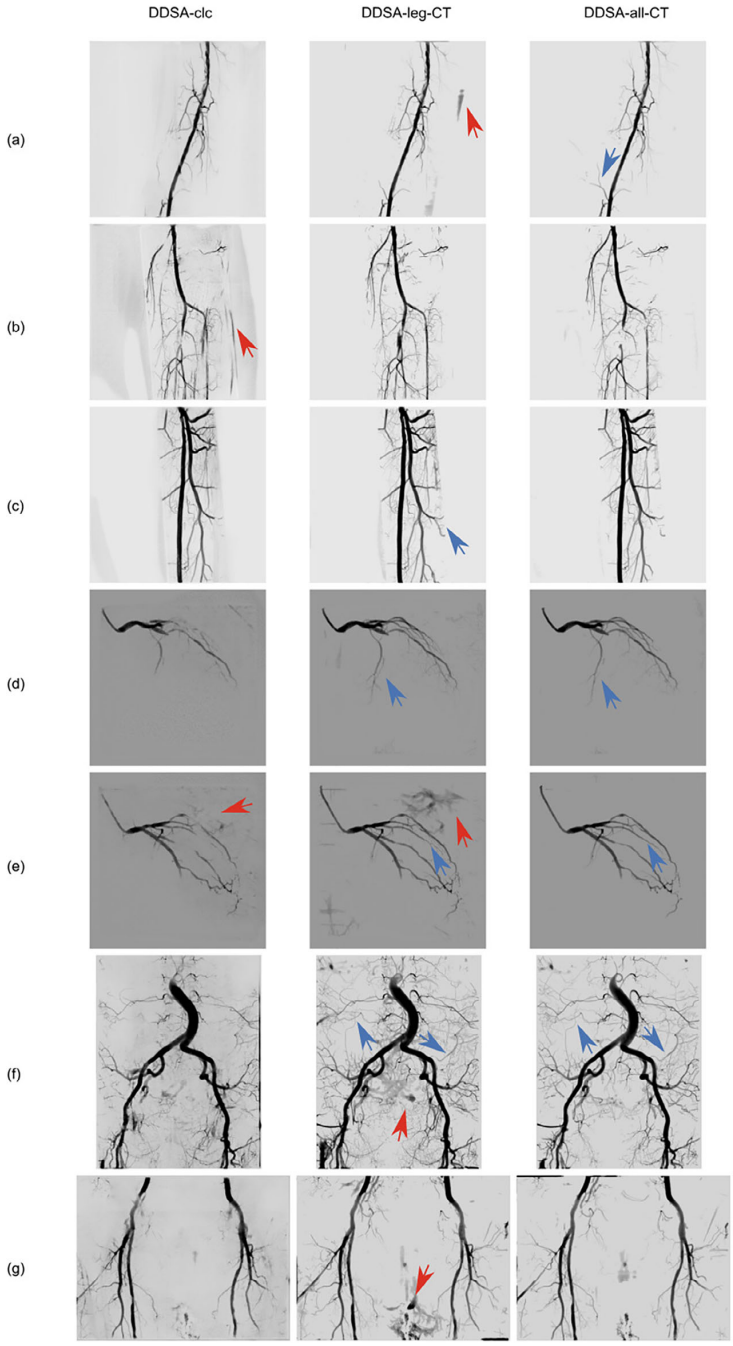


FIGURE B2 Results of different training set size of CT data: (a)–(g) represent different test data. Left to right: results of U-net trained on clinical data with 64 data (DDSA-clc-64), results of U-net trained on synthetic dataset with 25 CT data (DDSA-syn-25), results of U-net trained on synthetic data with 52 CT data (DDSA-syn-52), and results of U-net trained on synthetic data with 105 CT data (DDSA-syn-105). Blue arrows indicate more vessels extracted by DDSA models. Red arrows indicate misssing vessels and wrongly inferred vessels by DDSA models. DDSA, deep learning-based DSA.

TABLE B1 Quantitative results for analyzing the effect of contrast flow simulation: PSNR and SSIM with 95% confidence intervals.

	PSNR			SSIM		
Method	**Figure B1a**	**Figure B1b**	**Figure B1c**	**Figure B1a**	**Figure B1b**	**Figure B1c**
DDSA-non-flow	43.74 ± 1.30	35.06 ± 1.42	33.82 ± 5.15	0.9982 ± 0.0008	0.9804 ± 0.0027	0.9822 ± 0.0053
DDSA-flow	43.75 ± 1.49	36.32 ± 0.90	35.89 ± 1.62	0.9982 ± 0.0008	0.9820 ± 0.0026	0.9845 ± 0.0013

Abbreviations: DDSA, deep learning-based DSA; PSNR, peak signal-to-noise ratio; SSIM, structural similarity index matrix.

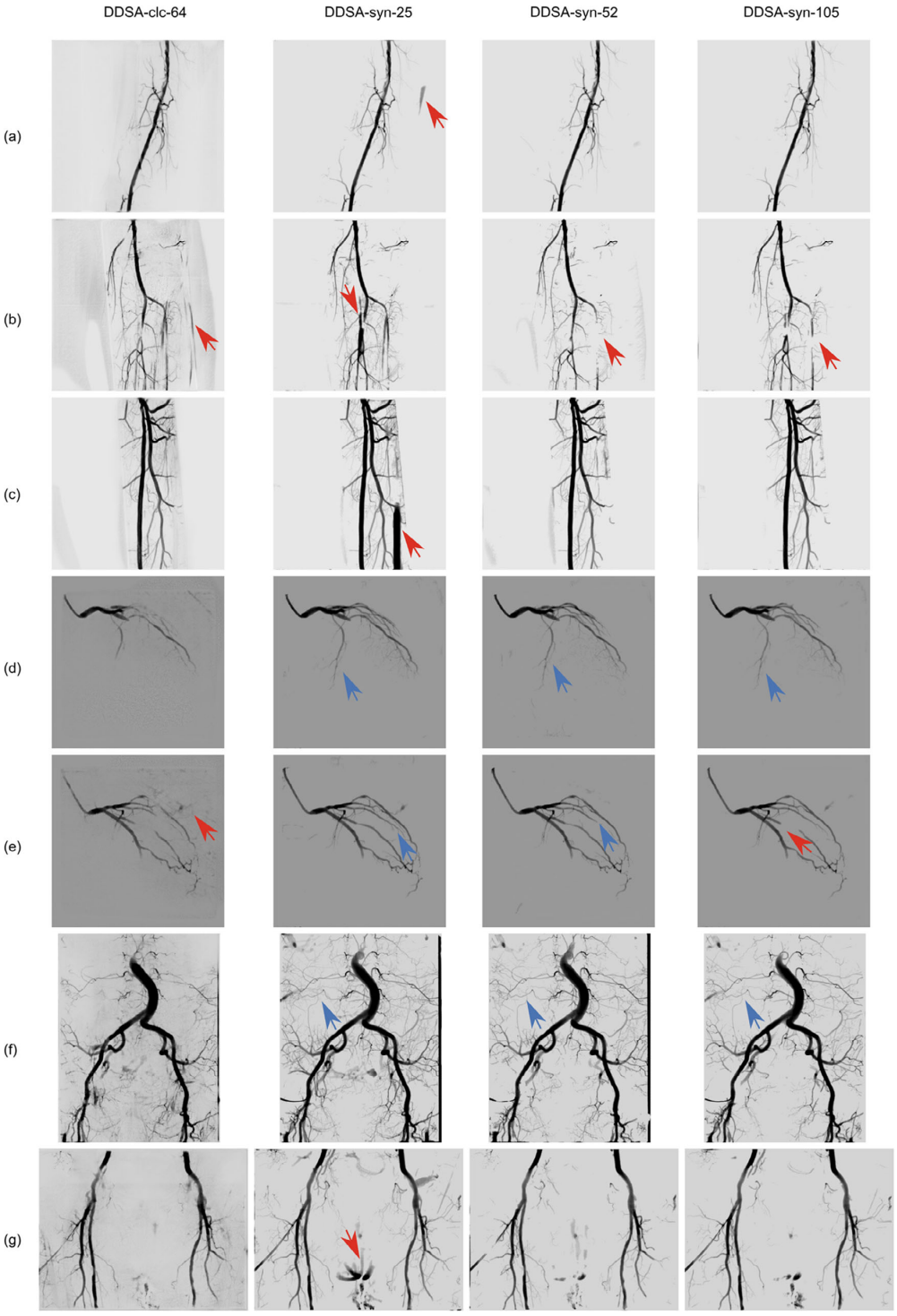


FIGURE B3 Results of using different CT anatomical sites: (a)–(g) represent different test data. Left to right: results of U-net trained on clinical data (DDSA-clc), results of U-net trained on synthetic data with leg CT data (DDSA-leg-CT), and results of U-net trained on synthetic data with all CT data (DDSA-all-CT). Blue arrows indicate more vessels extracted by DDSA models. Red arrows indicate wrongly inferred vessels by DDSA models.

TABLE B2 Quantitative results for analyzing the effect of the training set size of CT data: PSNR and SSIM with 95% confidence intervals.

Method	PSNR			SSIM		
	Figure B2a	Figure B2b	Figure B2c	Figure B2a	Figure B2b	Figure B2c
DDSA-clc-64	42.50 ± 1.97	34.86 ± 1.11	35.61 ± 3.30	0.9980 ± 0.0009	0.9780 ± 0.0018	0.9791 ± 0.0193
DDSA-syn-25	43.31 ± 1.19	34.74 ± 0.57	25.36 ± 0.59	0.9981 ± 0.0008	0.9804 ± 0.0027	0.9746 ± 0.0014
DDSA-syn-52	43.48 ± 1.40	35.45 ± 1.36	35.08 ± 2.14	0.9981 ± 0.0009	0.9840 ± 0.0027	0.9838 ± 0.0018
DDSA-syn-105	43.76 ± 1.48	35.53 ± 1.58	35.45 ± 2.01	0.9982 ± 0.0008	0.9810 ± 0.0029	0.9834 ± 0.0013

Abbreviations: DDSA, deep learning-based DSA; PSNR, peak signal-to-noise ratio; SSIM, structural similarity index matrix.

TABLE B3 Quantitative results for analyzing the effect of the CT anatomical sites: PSNR and SSIM with 95% confidence intervals.

Method	PSNR			SSIM		
	Figure B3a	Figure B3b	Figure B3c	Figure B3a	Figure B3b	Figure B3c
DDSA-clc	42.50 ± 1.97	34.86 ± 1.11	35.61 ± 3.30	0.9980 ± 0.0009	0.9780 ± 0.0018	0.9791 ± 0.0193
DDSA-leg-CT	42.95 ± 0.60	36.65 ± 0.69	36.08 ± 1.65	0.9979 ± 0.0008	0.9822 ± 0.0022	0.9835 ± 0.0014
DDSA-all-CT	43.75 ± 1.49	36.32 ± 0.90	35.89 ± 1.62	0.9982 ± 0.0008	0.9820 ± 0.0026	0.9845 ± 0.0013

Abbreviations: DDSA, deep learning-based DSA; PSNR, peak signal-to-noise ratio; SSIM, structural similarity index matrix.

APPENDIX C: ADDING POISSON NOISE TO CT FORWARD PROJECTIONS

We generated the noise using Poisson statistics associated with photons. The relation between the projection values p_t of a CT projection image P_t and the number of photons N_t measured at the detector is defined as

$$N_t = N_0 e^{-p_t} \tag{C1}$$

where N_0 is the initial number of photons. The correlated Poisson noise n_t was calculated with

$$n_t = G(P_{\text{ios}}(N_t) - N_t; \sigma_G) \tag{C2}$$

where $P_{\text{ios}}(x)$ represents the sampling results from a Poisson distribution with mean x, and $G(x, \sigma_G)$ means Gaussian smoothing operator to x with a standard deviation σ_G. The Gaussian smoothing was utilized to generate correlated noise and σ_G was a value randomly chosen between 0.3 and 1. Then, the number of photons was updated with

$$N'_t = \max(1, N_t + n_t) \tag{C3}$$

N'_t is restricted to strictly positive values to be converted to the corresponding x-ray projections p'_t. The final projection values with Poisson noise was calculated with

$$p'_t = -\ln\left(\frac{N'_t}{N_0}\right) \tag{C4}$$

Thus, the noise depends on the initial number of photons N_0.

We utilized the SNR as a metric to evaluate the noise level. The SNR ranges for both clinical fluoroscopy images and generated CT projected images are determined by calculating several homogeneous regions within these images. We adjusted the value of N_0 to ensure that the SNR range of the generated CT projection images closely matched the SNR range observed in the clinical images.

In our tests, the SNR values of the fluoroscopy images fall within the range of [45, 80]. To achieve this target SNR range, the initial number of photons N_0 was set to be uniformly sampled from the range of $[5 \times 10^4, 1.25 \times 10^5]$.