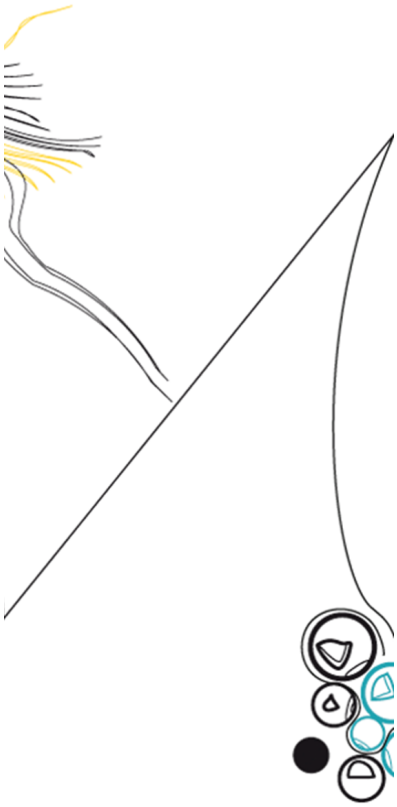
UNIVERSITY OF TWENTE.

Faculty of Science & Technology
Faculty of Electrical Engineering,
Mathematics & Computer Science
Data Science + Technical Medicine
Radiologie - LUMC

Representing Ultra Low Dose CT scans as Chest X-Rays: how far can and do we need to go?

Olivier Paalvast
s1590707
Master Thesis
September 2022



Supervisors:
Medical supervisors: prof. dr. H.J. Lamb
Technological supervisor: prof. dr. ir. C.H. Slump
Data Science supervisor: Dr. M. Poel
Daily supervisors: Dr. M. Sevenster
Dr. O. Hertgers
Dr. K.F.M. Hergaarden
Process supervisor: Dr. R. M. Krol

University of Twente
7500 AE Enschede
The Netherlands

CONTENTS

Introduction

I. Motivation

This thesis focuses on the idea of replacing conventional chest radiography with Ultra Low-Dose Computed Tomography (ULDCT) imaging. For chest imaging specifically, the clinical value of ULDCT over the traditional Chest Radiograph (CXR) has been shown in multiple clinical areas. The primary objector to the use of ULDCT is the clinical interpretation time, which is on average ten times as long as the CXR interpretation time. This difference makes the widespread adoption of ULDCT imaging unfeasible at this point. Resolving this issue entirely goes beyond the scope of this thesis project. To direct my efforts, I have chosen to focus a number of promising methods related to the interpretation of ULDCT data.

The goal of this thesis is therefore to investigate and develop methods which may aid in the interpretation of ULDCT imaging. One promising approach is to use Digitally Reconstructed Radiograph (DRR)s to provide a synthetic CXR of ULDCT data. The intuition here is that CXRs are faster to read, so therefore a DRR might be too. There has, however, been very limited research into the diagnostic value of DRRs.

In this thesis I have investigated the manner in which DRRs can be used to create representative visualisations of ULDCT data. The goal in this has always been to approach the quality of the original CXR, not to match or surpass it. The DRR could be an ideal vessel to display summary information of an ULDCT scan in a format that is highly familiar to radiologists.

II. Background

The use of X-ray imaging in medical diagnostics is highly prevalent with the Chest Radiograph (CXR), or the X-ray of the thorax, generally being the first diagnostic examination applied when pathologies of the chest are suspected [1–3]. The CXR is easy and relatively cheap to produce [4], comes with a minimal radiation exposure to the patient [5] and is quick to interpret for radiologists ($\pm$ 1.5 minutes) [6, 7]. Despite these advantages, the diagnostic value of CXR suffers from potential tissue homogeneity and the superimposition of tissues in the thorax when the radiograph is taken, which could occlude lesions and lead to missed diagnoses [7]. Computed Tomography (CT) is able to provide a more detailed volumetric visualisation of the thoracic structures and is well-established in thoracic imaging, but comes with a significantly higher radiation exposure for the patient [8], a longer scan time and increased cost [7] and significantly increased interpretation time for radiologists ($\pm$ 15 minutes) [6]. The effective radiation dosage for a CXR is 0.10 mSv (range: 0.01 - 0.26 mSv) compared to 5.5 mSv (range: 2.0 - 20.4 mSv) for a chest CT, which is associated with a cancer risk of 1:2000 [5, 8].

Advances in CT scanners and reconstruction methods have enabled the creation of Low-Dose (LD) and Ultra Low-Dose (ULD) CT scans [2]. For chest examinations, Low-Dose Computed Tomography (LDCT) scans are associated with an effective radiation dose of 2 mSv (range: 1.5 - 2.5 mSv) [9, 10], whereas Ultra Low-Dose Computed Tomography (ULDCT) scans are associated with an effective dosage comparable to a CXR (range: 0.07 - 0.27 mSv) [1, 11]. The diagnostic value of LDCT has been extensively proven for lung cancer screening, whereas the sensitivity of CXR in patients with lung cancer symptoms was shown to be only 77-80% [12]. In 2011, the National Lung Screening Trial showed a relative reduction in mortality of 20% with LDCT screening compared to CXR screening after a median follow-up of 6.5 years [13]. The value of LDCT lung cancer screening was further corroborated in 2020 by the NELSON study, finding a cumulative rate ratio for mortality of 0.76 in the LDCT screening group compared to the no screening group [14]. A 2021 review [2] found LDCT and ULDCT to have high diagnostic accuracy for honeycombing and bronchiectasis and pneumothorax, consolidations and ground glass opacities respectively.

The primary objector to the widespread use of LDCT over CXR despite overwhelming clinical evidence is the clinical interpretation time. Cowan et al.[6] showed that CT scans on average take ten times more time for interpretation compared to CXR (15 minutes versus 1.5 minutes). This is a significant enough difference in interpretation time that it is (financially) unfeasible to replace a majority of CXR examinations with LDCT without overwhelming an already busy radiologist workflow. In order to cut down on the clinical interpretation time of diagnostic imaging several methods have been proposed. One of these methods is the use of Computed-Aided Diagnostics (CAD) methods. Over the past few years Deep Learning (DL) based applications have dominated this field, showing success in the detection of lung cancer [15], pneumonia [16], tuberculosis [17] and recently COVID-19 [18]. These methods generally focus on CXR and axial reconstruction slices of CT data which are then used for segmentation, classification or detection using Convolutional

Neural Network (CNN) based architectures [19]. These networks are in part as successful as they are due to their ability to work with morphological information without having to pre-define features specific to their task [20]. The acceptance of DL-based CAD systems in radiology workflows is on a slow but steady rise, as a 2019 survey [21] found a majority of radiologists in favour of their use.

Another method that could be employed to reduce clinical interpretation time is the use of so-called Digitally Reconstructed Radiograph (DRR) [22, 23]. A DRR is a reconstruction of summations over simulated projection lines through volumetric imaging data. These reconstructions are generally used in image registration in radiotherapy [24–27] but could also see use in diagnostics as the traditional Posteroanterior (PA) and lateral CXRs can be reconstructed from chest CT data [28–31]. A DRR shares in the advantage of CXR in that it is quick to interpret superficially. Additionally, it could be used to guide a radiologist more specifically through the volumetric (ULD)CT data it was constructed from. A segmentation of the CT data can for example be used to calculate affected lung volume in COVID-19 patients which can then be projected onto the CXR reconstruction for quick interpretation [31, 32]. Limiting factors in the use of DRRs are resolution, which is limited by the slice thickness of the CT data its reconstructed from, and image quality, which depends both on the quality of the CT data as well as the DRR reconstruction algorithm used.

III. Research Questions

Realising the replacement of CXR imaging with ULDCT imaging is a task that goes well beyond the scope of only a master thesis. This means choices have to be made with regards to what this master thesis focuses on. This focus is to investigate which methods exist to generate and optimise CXR-like representations of ULDCT data. This ties into the goal of reducing the clinical interpretation time by displaying key information in a singular image. Additionally, this is displayed in a format well known to radiologists.

From this focus I've defined the following four research questions:

1) *What methods exist to generate synthetic chest X-Rays from (ULD)CT data and how are these perceived quantitatively and by clinical experts?*
2) *Can AI-models trained for chest X-Ray disease classification be used to evaluate the (diagnostic) image quality of Digitally Reconstructed Radiographs?*
3) *Can AI models be used to generate realistic CXR and can they subsequently facilitate the generation of a CXR visualisation for a DRR?*
4) *To what extent can super resolution models boost the perceived quality of DRRs constructed from ULDCT data?*

My contributions in this thesis can be summarised by the following:

1) An evaluation of existing DRR generation methods with clinical experts.
2) A cross-domain application of State-of-the-Art image classification models to CXRs and DRRs.
3) A model capable of generating realistic CXRs and optimising a CXR-like representation of a DRR.
4) An evaluation with clinical experts of the application of a State-of-the-Art Super Resolution model to CXRs and DRRs.

IV. Document structure

This thesis is organised as follows. There are four main chapters, each tackling one of the research questions. These chapters are written to be readable as a stand-alone chapter. This means that each has an introduction, methods, results, discussion and conclusion section specific to that chapter. After these chapters a general discussion is included in which the findings of each of the chapters are combined and discussed. This preempts the conclusion which ties the document together.

Creating synthetic chest X-Rays from ULDCT data

I. Introduction

A Digitally Reconstructed Radiograph (DRR) is a reconstruction of summations over simulated projection lines through volumetric imaging data. By tracing virtual X-Rays through a volume, and by accounting for the attenuation that would otherwise occur, a virtual reconstruction is obtained which resembles a conventional radiograph. These reconstructions are generally used in image registration in radiotherapy [24–27] but could also see use in diagnostics as the traditional Posteroanterior (PA) and lateral Chest Radiograph (CXR)s can be reconstructed from chest CT data [28–31]. Examples of DRRs being used as a diagnostic tool include the quantification of emphysema [33], the inspection of flatfoot deformity [34] and the automated quantification of covid infection spread [31].

A DRR shares in the advantage of the CXR in that it is quick to interpret superficially and that its format is very well known to radiologists and medical experts. Additionally, it could be used to guide a radiologist more specifically through the volumetric (ULD)CT data it was constructed from. A segmentation of the CT data can for example be used to calculate affected lung volume in COVID-19 patients which can then be projected onto the CXR reconstruction for quick interpretation [31, 32]. Given these benefits, there is a potential role for drrs in the reduction of the clinical interpretation time for the (ULD)CT scans they're constructed from [22, 23].

Even though DRRs can and have been used as a diagnostic tool, there has not yet been an overarching clinical evaluation comparing the underlying mechanisms with which they are constructed [29, 31, 33–35]. Carey et al. [35] clinically evaluated their proposed DRR construction mechanism, but did not compare it to other construction mechanisms. In the work of Zhang et al. [31], an infection-aware DRR was proposed with an adjustable amount of radiological signs of infection. This novel approach was limited primarily in the highly specific and evolving disease pattern on which it focused. Moore et al. [28] evaluated their proposed drr construction method with clinical experts, though they themselves noted that their research was limited by their optimisation for a specific CT acquisition system.

The need for a clinical evaluation of DRR construction methods has led to the following research question:

> *What methods exist to generate synthetic chest X-Rays from (ULD)CT data and how are these perceived quantitatively and by clinical experts?*

To answer this question, this chapter is structured in the following way. In section II the basics of X-Rays, CT imaging and DRR generation are discussed. A literature review is presented in section III from which key DRR generation methods are identified. In section IV an automated histogram-based analysis and a clinical reader study are proposed. These are reported on in section V and discussed in section VI. A conclusion is provided in section VII.

II. Background

This background section seeks to inform readers of specific speciality backgrounds of core principles in other fields.

A. Conventional X-Ray imaging

X-Rays are a form of high-energetic electromagnetic radiation that can penetrate human tissue. The invention and use of X-Rays in medical practice have been closely linked since their inception. Mere weeks following the 1895 submission of Wilhelm Roentgen of his paper on the discovery of X-Rays, a clinical application was developed to image a needle stuck in a hand. A month later the technique was applied during a surgical operation [36]. Roentgen went on to win the first Nobel prize in physics for his invention. Conventional radiography has changed much since the time of Roentgen, but several core principles survive to this day. Diagnostic X-Ray setups still use an X-Ray source and a detector to image a patient.

1) The X-Ray source: The X-Ray source, commonly referred to as the tube, consists of a positive and a negative electrode; the anode and cathode respectively. The anode and the cathode are encapsulated in a vacuum. A schematic representation of the X-Ray source is included in figure 1. The cathode, which is usually made of tungsten, emits electrons when heated. These electrons are then accelerated towards the anode with a certain acceleration potential; the tube voltage. By definition, the kinetic energy (in electron Volt (eV)) of the accelerated electron is equal to the potential which it has been accelerated by, such that a tube voltage of 100 kV results in a kinetic energy of 100 keV for the electron. The anode, which is also made out of tungsten, is then bombarded by the electrons. Here two processes occur. A large part of the electrons will undergo characteristic interactions with the atoms of the anode through ionisations and excitations. This energy is dissipated as heat.

A much smaller part of the electrons (roughly 1%) will instead be decelerated as they pass by the atomic nuclei of

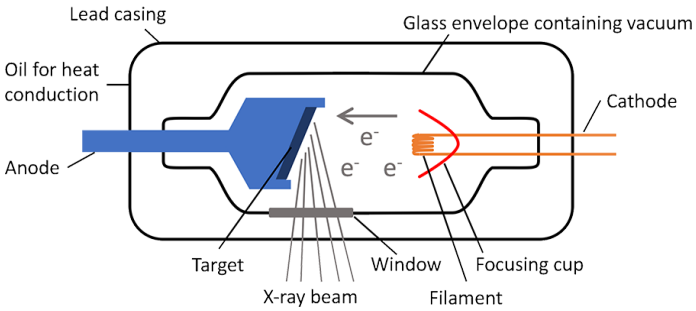


Fig. 1: Schematic representation of an X-Ray source. The X-Ray source, or the tube, consists of an anode and a cathode encapsulated in a vacuum. The cathode, when heated, emits electrons which are accelerated towards the anode by the tube voltage. Upon impact a large part of the kinetic energy of the impacting electrons is dissipated as heat. Roughly 1% is converted into Bremsstrahlung and emitted through a filtering window that stops low-energetic particles. Image sourced from [37]

the anode. The subsequent loss of kinetic energy is converted into a photon which is then emitted as radiation; the 'X-Ray' or so-called Bremsstrahlung. This Bremsstrahlung has a continuous spectrum, as shown in figure 2, which is related to the energy of the impacting electrons. The peaks in the spectrum are the characteristic K-shell photons of the anode material. The emitted X-Rays pass through a window which filters out low-energetic particles.

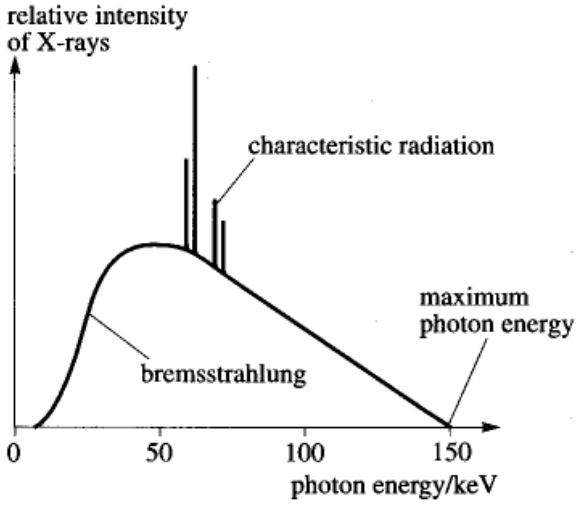


Fig. 2: The energy spectrum of Bremsstrahlung. Shown here is the relative intensity of X-Rays at specific photon energies. The peaks in the spectrum correspond to the characteristic K-shell photons of the anode material. Image sourced from [37]

2) The interaction of X-Rays with matter: As the X-Rays exit the tube and hit the patient interactions start to occur between the matter of the patient and the impacting X-Rays. These interactions are essential as it is the variation in the transmission of photons through the patient that gives rise to the X-Ray image. The X-Rays interact through either photoelectric absorption or scattering, which is then divided into incoherent Compton scattering and coherent scattering. If an X-Ray photon undergoes such an interaction it is considered lost to the primary radiation. The rate at which photons are lost is proportional to the thickness of the medium (dx) it passes through, as well as the number of incident photons (N) and is given by:

$$dN = -\mu dx \tag{1}$$

where μ is the Linear Attenuation Coefficient (LAC). This describes the probability (p) per unit length (x) for an X-Ray photon of certain energy to interact when passing through a medium:

$$\mu = \frac{dp}{dx} \tag{2}$$

The X-Ray photons are attenuated according to the following equation:

$$N(x) = N(0) \cdot e^{-\mu x} \tag{3}$$

This shows that X-Ray photons are attenuated exponentially as their depth in the medium increases. The tube potential plays a role here, as the LAC is smaller for an X-Ray photon of high energy.

Of the possible interactions, the photoelectric effect describes the process in which the X-Ray photon is absorbed and a photo-electron is emitted. The probability of this is practically inversely proportional to the energy of the X-Ray photon. Incoherent Compton scattering occurs whenever an X-Ray photon collides with an atomic electron. In this process the X-Ray photon is scattered, i.e. its direction is altered, and it continues with reduced energy. The difference in kinetic energy is preserved through the release of a photon. The higher the X-Ray photon energy, the more likely it is to be scattered in a forward direction. In coherent scattering an interaction with an atomic electron does not transfer energy and only the direction of the X-Ray photon is altered. The probability of coherent scattering is inversely related to the X-Ray photon energy. The probability of such interactions in a patient depends on the atomic number of the matter that is interacted with. In humans, calcium has the highest atomic number, which means that interactions are more likely to occur in the denser bone regions than for example the lungs.

The contrast, as part of the quality of an X-Ray image, is determined by the object thickness and the energy spectrum of the impacting photons. Without considering photon interaction effects, this already requires knowledge regarding the fraction of photons that make it to the detector. This depends on the physical characteristics of a patient. When the interactions such as scattering are considered on top of this, the contrast is degraded. Efforts to minimise photon interactions help in optimising image quality, such as increasing the photon energies by increasing the tube voltage. This is not always feasible as it can also increase the malignant effects to the patient.

3) The X-Ray detector: The basis of an X-Ray detector is a substance or device which can record the impact of X-Rays. The impact darkens the X-Ray image, such that areas that let through a large amount of X-Rays with relatively little interaction, such as the lungs, obtain a darker shade. Areas that do have a lot of interaction, such as bones, are coloured white. The field of X-Ray detectors had seen relatively little development since the inception of the X-Ray as until two decades ago the films used were conceptually the same as the ones Roentgen used originally. The concept of the film revolved around the idea of creating a permanent and fixed recording of X-Ray imaging. To achieve this, a transparent film was typically coated with silver bromide. When struck with X-Rays the coating would absorb the energy and would, when developed, be reduced to metallic silver specks. The resulting film would then absorb visible light wherever ionising radiation struck. To reduce the effect of scattering so-called Bucky grids are placed over the detector [38].

In the past two decades digital detectors have largely replaced film based detectors in medical imaging. Digital detectors generally permit recording of images with up to 400 times the dynamic range when compared to film [38]. Advances in computer storage capabilities had made it feasible to make this transition, as the digital availability of X-Ray images greatly speeds up retrieval, exchange and copying as well as post processing to, for example, apply certain window levels.

4) The Chest Radiograph: The Chest Radiograph (CXR) is an extremely commonly used medical diagnostic tool. The goal is to display at least the entirety of the lungs, from base to apex such that the pleural cavities can be examined. A CXR can be created in one of two ways. The Posteroanterior (PA) X-Ray is created when the patient has his/her chest facing the detector. As can be seen in figure 3, the effect of the diverging beam on the size at which the heart is displayed is limited. Whenever a reference is made to a CXR, it is generally a PA X-Ray that is being referred to.

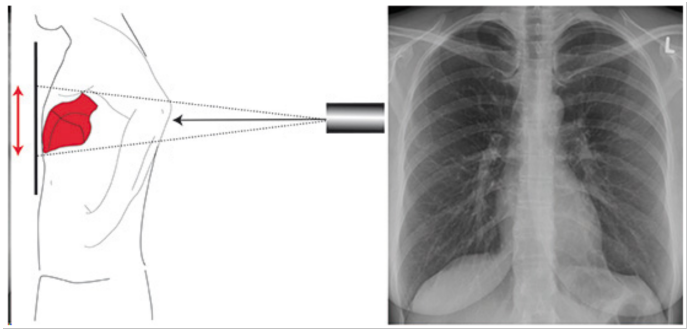


Fig. 3: Schematic representation of the creation of a Posteroanterior (PA) X-Ray. This X-Ray is generally taken standing up with hands placed on the hips. Because the heart lies distally in the diverging X-Ray beam, the effect of enlargement on the heart is limited. Adapted from [39].

The alternative, the Anteroposterior (AP) X-Ray, is generally made whenever a patient is unable to stand. Here the patient has their back to the detector, which would commonly occur when the image is created bedside or sitting down. As can be seen in figure 4, this causes the heart to appear larger on the detector than it would for the PA X-Ray.

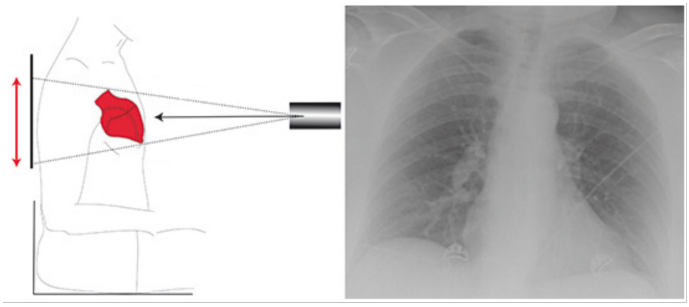


Fig. 4: Schematic representation of the creation of a Anteroposterior (AP) X-Ray. This X-Ray is generally taken either in seated or lying position. Here the heart lies proximal in the X-Ray beam and as such is enlarged by the diverging X-Rays on the detector. Adapted from [39].

B. CT imaging

The CXR is an incredibly useful tool in medical diagnostics, but it suffers from being the limitation of being a 2D image of a 3D patient. The superposition of tissue that occurs when the 3D body is reduced to a 2D image could occlude valuable information. In the 1970s a new diagnostic tool was developed to circumvent this issue; the CT scanner. Computed Tomography (CT) is used to create cross-sectional (tomographic) images, or slices, of the human body. This is achieved by using a source and detector housed in a gantry that can move around a patient, as is shown schematically in figure 5. Originally, a CT scanner would make one full revolution after which the patient had to be moved to image another slice. Modern CT scanners create a continuous helical or spiral image as the patients' bed moves through the gantry.

As a CT scanner also makes use of X-Rays, many of the principles that apply to conventional X-Ray imaging apply here as well. At its core a CT scanner measures the attenuation of X-Rays through human tissue. A key difference between conventional X-Ray imaging and CT imaging is the way the detector is set up. Modern CT scanners use an array of 64, 128 or even more detectors. These CT detectors do not directly produce an image. Instead, they measure the attenuation of the specific part of the X-Ray beam that is aimed at them. This process is repeated as the gantry revolves around the patient.

1) Reconstruction algorithms: To obtain a characteristic CT slice, the raw CT data has to be reconstructed into an image. As the gantry revolves around the patient many projections of the body are recorded. These projections can be combined to reconstruct the original image in a process called simple back-projection. An example of this is included in

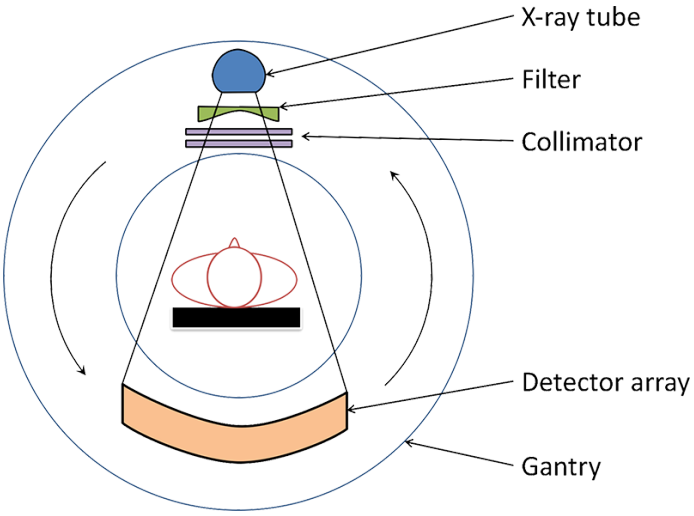


Fig. 5: Schematic representation of a CT scanner. Much like the X-Ray setup, there is an X-Ray source as well as a detector. In a CT scanner these are housed in the gantry, which allows them to revolve around a patient. The bed upon which a patient is placed can move through the gantry. Image from [40].

figure 6. As more views are used to compute the final image, the accuracy of the representation increases. In addition to using more views, the individual views can be filtered using a specific window to increase the accuracy of the representation at its boundaries. This is called Filtered Back-Projection (FBP).

Further developments in reconstruction algorithms have centered around iterative reconstruction. Here FBP is used to create a primary image of the raw data. This is then compared to the raw data such that an improved and updated image can be generated. This iterative process is repeated until a preset value is obtained. Model-based iterative reconstruction is a further improvement where statistical measurements and modeling of the CT scanner are taken into account in the reconstruction process.

The pixels in the reconstructed CT slices are scaled with Hounsfield Units (HU). This is a scale that is used to indicate relative densities. By design, Hounsfield chose air to have a value of -1000 HU, fat -60 to -120 HU, water 0 HU and bone +1000 HU. Using the HU values, certain window levels can be applied to enhance contrast in for example the bones, lungs or soft tissue.

2) Radiation exposure: The radiation dose the patient receives during a CT scan is quantified using the effective dose, which is measured in milliSieverts (mSv). Factors such as the scan time, the pitch and the size of the patient all play a large role in determining the effective dose. In addition to this, the tube current and tube voltage determine how many X-Rays hit the patient and how energetic these are. A higher energetic X-Ray is able to undergo more interactions in the patient and can therefore inflict more damage. This damage

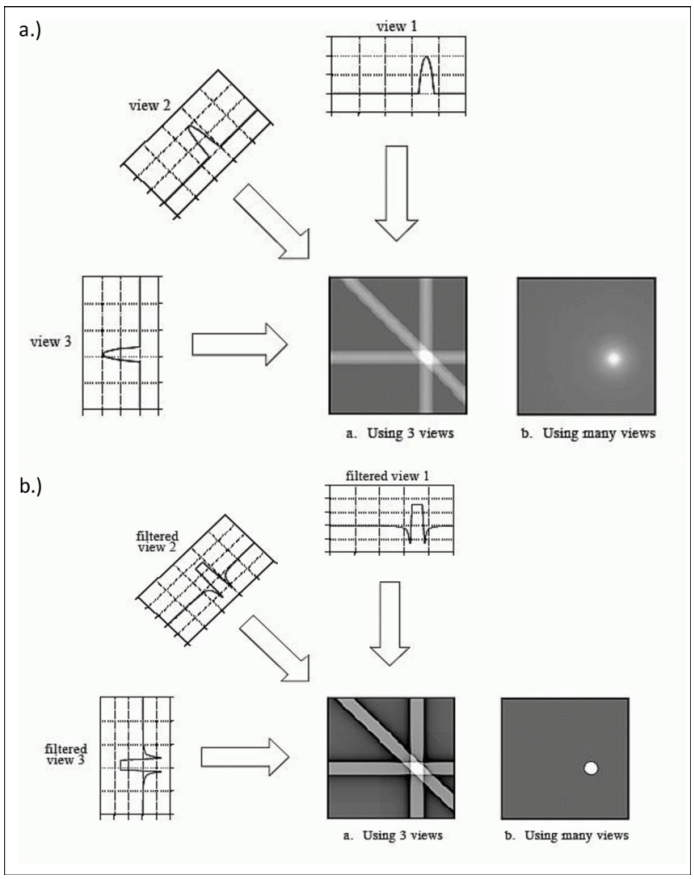


Fig. 6: a) Schematic representation of the simple back-projection algorithm. The recorded views from different angles are combined to form an image. The more views are included the more accurate the representation becomes. b) Schematic representation of filtered back-projection. In filtered back-projections the individual views are filtered with a specific window to increase the accuracy of the representations of the boundaries of imaged objects. Image from [37].

manifests at a cellular level in the destruction of parts of the DNA. To keep this damage to a minimum, the principle of As Low As Reasonably Possible (ALARP) is applied to diagnostic imaging. This principle tries to seek a balance between the level of radiation exposure on the one hand and the diagnostic image quality required to accurately detect pathologies.

In CT the effective dose to a patient can, amongst things, be controlled by lowering the tube current. When referring to ULDCT, the Ultra Low-Dose is achieved by reducing the tube current to a minimum of 10 mA. The consequence of this reduction is a degradation in the contrast and therefore the discriminating ability of the CT images. One of the drivers behind the development of CT scanners and reconstruction algorithms has been the effort to reduce the radiation exposure to the patient, whilst retaining diagnostic imaging quality. Using more advanced reconstruction techniques a greater level of noise can be removed at ever decreasing amounts of radiation exposure.

C. Digitally Reconstructed Radiographs

A Digitally Reconstructed Radiograph (DRR) is a synthetic X-Ray image that has been simulated by digitally tracing X-Rays through a 3D CT volume. The traced Radiological Path Length (RPL) of an X-Ray enables the calculation of the attenuation of that ray with regards to the tissue it passed. If this is repeated often enough for enough entry points a simulated X-Ray image can be obtained. In the construction of a DRR the Hounsfield Units (HU) of the CT scan have to be converted back into the Linear Attenuation Coefficient (LAC) of an X-ray image. In the traditional X-ray image, the pixels represent the attenuation of the X-ray beam from the source to the detector. In a DRR a pixel is calculated in a similar fashion. Instead of an X-ray source, a simulated ray caster is used. From this source, such as a point source, X-ray beams are virtually cast through the CT volume in a process highly similar to taking an X-ray. A schematic overview of this is shown in figure 7.

For every virtual ray that is cast, the intersection of that ray with the voxels of the CT volume is calculated. The RPL, or the total distance a ray travels through the CT volume, is used in combination with the HU values of the intersected voxels to calculate an attenuation coefficient for a specific pixel in the DRR. This relationship can be described using the Beer-Lambert law [41]:

$$I = I_0 e^{-\mu x} \tag{4}$$

Where I_0 is the incident beam, x the distance travelled, I the intensity of the beam after travelling distance x and μ the LAC. For a parallel projection the average LAC can be computed:

$$\mu_{av}(x,z) = \sum_{y=1}^{N} \frac{\mu_{water}(C(x,y,z)+1024)}{N \cdot 1024} \tag{5}$$

where $C(x,y,z)$ represents the CT volume [42, 43], and the 1024 is used to compensate for the HU scale. This equation can be used in combination with equation 4 to compute the DRR:

$$I_{DRR}(x,z) = e^{\beta \cdot -\mu_{av}(x,z)} \tag{6}$$

Here β is a parameter that regulates the relationship between I and the HU in the volume data. By repeating this process for all rays that hit the volume a DRR reconstruction can be created.

The construction methods for DRRs can broadly be sorted into two categories: point-source based projections, as shown schematically in figure 7, and parallel based projections, as shown schematically in figure 9. Point-source based projection methods more closely mimic the actual construction of an X-Ray as it introduces a level of divergence to the image. This is shown schematically in figures 4 and 3. Parallel based projections forego this divergence by arguing that at significant distance from the detector to the source, the x-ray beams are near parallel.

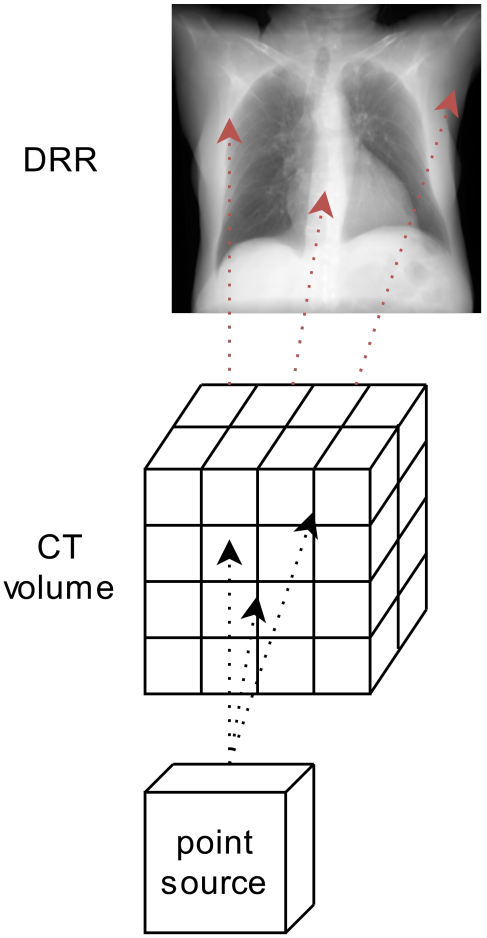


Fig. 7: Schematic representation of the creation of a DRR from a point source. The simulated X-Rays are shown in black before they hit the CT volume and in red after. The attenuation of the simulated X-Rays through the volume is summed to obtain a pixel in the DRR. Because of the point source nature, a certain level of divergence is seen in this DRR.

For DRRs constructed from standard CT data this argument holds up. But in many applications where the CT data also consists of a divergent beam, such as in cone-beam CT, this doesn't apply. Cone-beam CTs are used frequently in an intraoperative setting which is where a lot of DRRs were originally made. For this reason, a lot of effort has been placed into efficiently computing a point-source based DRR.

The challenge in computing a point source DRR is at its core a ray-tracing problem. Rays, or in this case virtual x-rays, have to be traced from the point source through the CT volume to the detector. To calculate the pixel value at the detector, the Radiological Path Length through each voxel has to be calculated. This has to be repeated for each pixel in the detector plane, making this a computationally expensive operation ($O(n^3)$). A schematic representation of this challenge is shown in figure 8.

Siddon et al. [22] were the first to redefine this problem to be able to solve it in a more efficient manner. They approached

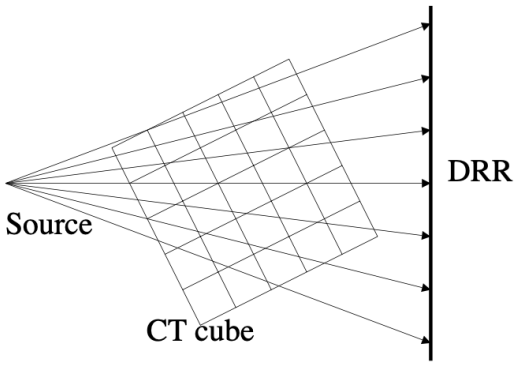


Fig. 8: Simplified schematic representation of the challenge in calculating the RPL through voxels from a point source. By originating from a point source every ray has a unique path through the volume for which a RPL has to be computed. Image from [44].

the CT data from a perspective of an intersection of three orthogonal planes instead of a collection of voxels. This was later improved upon by Jacobs et al. [23] by calculating the entry and exit points of rays through voxels more efficiently. Their work was done in a time when GPU processing power was not readily available. Nowadays a CT volume can easily fit into computer memory with a 300 slice CT scan taking up approximately 500 MB of memory. Nevertheless, computational efficiency is still a goal to strive for as non-parallelised operations will place a burden upon a CPU. The parallelisation offered by GPUs has greatly increased the speed at which a DRR can be computed and has enabled the development of further applications [45–47]. Rapid DRR computation can, for example, improve intraoperative patient registration [45].

The alternative parallel source based DRR projection method is computed by tracing an x-ray orthogonally through a CT volume. A schematic representation of this is given in figure 9. In this approach the RPL does not have to be computed as per the orthogonality of the virtual x-ray the RPL through each voxel is 1. As a result, this makes the computation of the DRR far easier and therefore faster than the point source based computation method. The parallel projection method comes at the cost of sacrificing the projection set-up accuracy with regards to realistically mimicking the set-up as it is performed in a regular CXR. The difference in resulting images is shown in the overview of figure 12.

III. Related work

This section discusses related work on DRRs with respect to their construction methods and their clinical application.

A. Construction methods and virtual interactions

The method with which a DRR is constructed can greatly influence the (diagnostic) image quality and potential

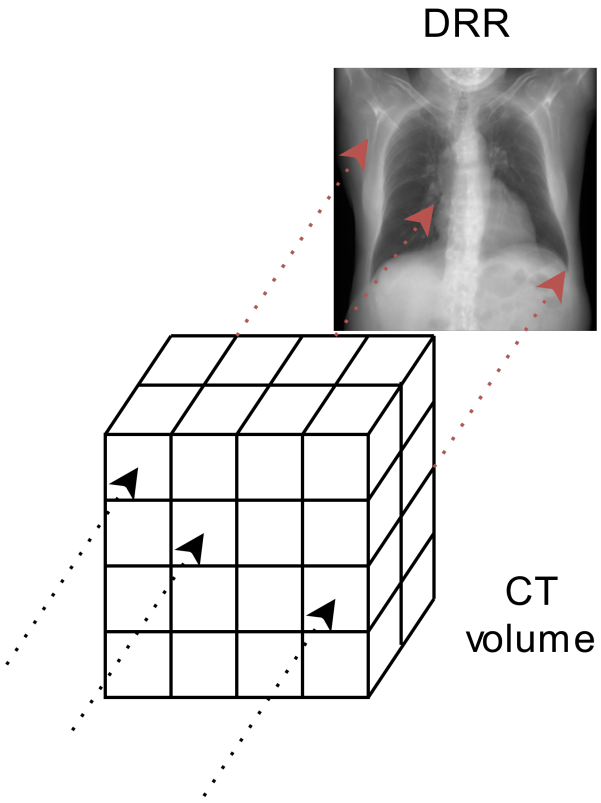


Fig. 9: Schematic representation of the creation of a DRR from a parallel projection perspective. The simulated X-Rays are shown in black before they hit the CT volume and in red after. Each simulated X-Ray only interacts with voxels in a straight line running from the front to the back of the volume.

resemblance of the resulting DRR to a conventional radiograph of the same domain. In all work on this topic a relationship is described between the interaction of simulated X-Rays with the virtual 3D CT volume. In both the simulation of virtual X-Rays as well as the interaction of said rays with the 3D data key differences pop up.

For point source based projections the underlying mechanism is largely the same across all related work as they all describe ray tracing from a virtual point source through a virtual volume to a virtual detector. This ray tracing is fundamentally based on the work by Siddon et al. [22]. One work that goes a step further is the work by Unberath et al. [30]. In their 'DeepDRR' approach, the authors approach the calculation of attenuation from a perspective of material decomposition for the interaction of simulated X-Rays. A segmentation of the CT volume is obtained to then assign a tissue-based weighing function to the interacting virtual X-Rays. By distinguishing between bone, soft tissue and air the authors propose that the resulting DRR more accurately mimics reality. Furthermore, scatter estimation is performed to add additional noise into the resulting image.

On the side of the parallel based projections multiple

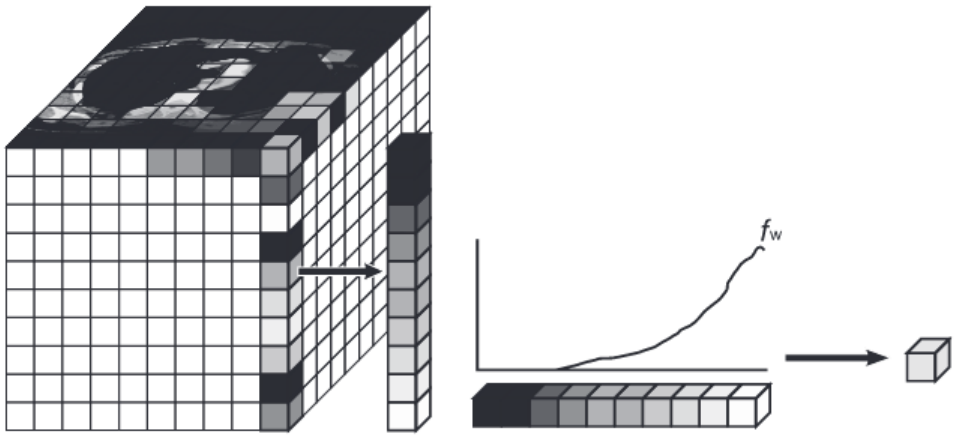


Fig. 10: Schematic representation of the voxel sorting approach used in the 'softMip' parallel based DRR construction method. In this approach voxels are sorted by value in a parallel line running sagitally, coronally or axially depending on the desired resulting image. Image from [48].

ideas regarding the interaction between virtual X-Rays and the 3D data exist. Campo et al. [33] apply the Lambert-Beer law [41] to compute the attenuation used in the construction of the DRR, see also subsection II-C. By applying this transformation an absolute weight is assigned to every voxel based on its HU value. Due to the exponential relationship in equation 4 this assigns an exponentially increasing value to dense, i.e. ossal, structures.

Meyer et al. [48] follow a different philosophy. Instead of assigning an absolute weighing factor to a certain HU value, they propose sorting voxels based on their HU value and then assigning a fixed weight based on the sorted position. A schematic overview of this sorting process is shown in figure 10.

The sorted voxel weighing factor is then described by the following relationship:

$$f_w^{softMip}(x) = \begin{cases} \frac{x}{2}; & \text{if } \mathrm{x} \leq 50 \\ 1.5x - 0.5; & \text{else} \end{cases} \tag{7}$$

where x is the absolute position between 0 and 100 of the voxel in the sorted array. Carey et al. [35] described a similar process of voxel sorting and assigning weights. In their work tomographic slabs were created based on a custom sorted voxel weighing factor relationship. This relationship was optimised visually, resulting in a 'wedge' that described the weighing factor for each sorted voxel position respectively. Altering this relationship can substantially alter a DRR, as is shown in figure 11. Here a moderate alteration in the voxel weighing relationship completely alters the resulting image. In figure 12 an example of the DRR construction methods described by Unberath et al. [30], Campo et al. [33], Meyer et al. [48] and Carey et al. [35] is shown.

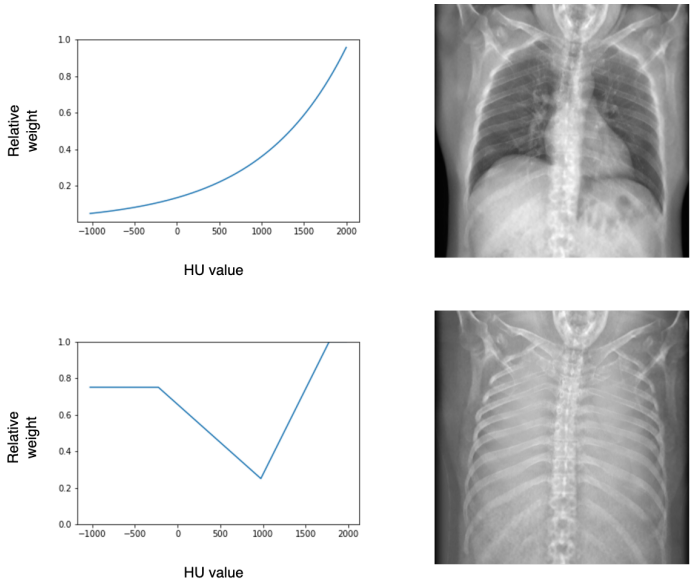


Fig. 11: Two example DRRs constructed for the same patient case. The first column the relationship between the HU value of a voxel and its weighing factor in construction the DRR. The first row shows a DRR as described by Campo et al. [33]. The second row shows a custom DRR which displays the effect of altering the DRR construction method on the resulting DRR.

B. Clinical applications of DRRs

Depending on the clinical application of DRRs, a point source based or parallel based projection method is applied. Notable examples of the use of point source based DRRs include 2D-3D image registration [49], DRR to portal image registration in radiotherapy [50], image registration in image guided interventions [46] and fluoroscopy guided procedures [30]. In these applications authors discuss the importance of fast computation with acceptable image quality, because time is

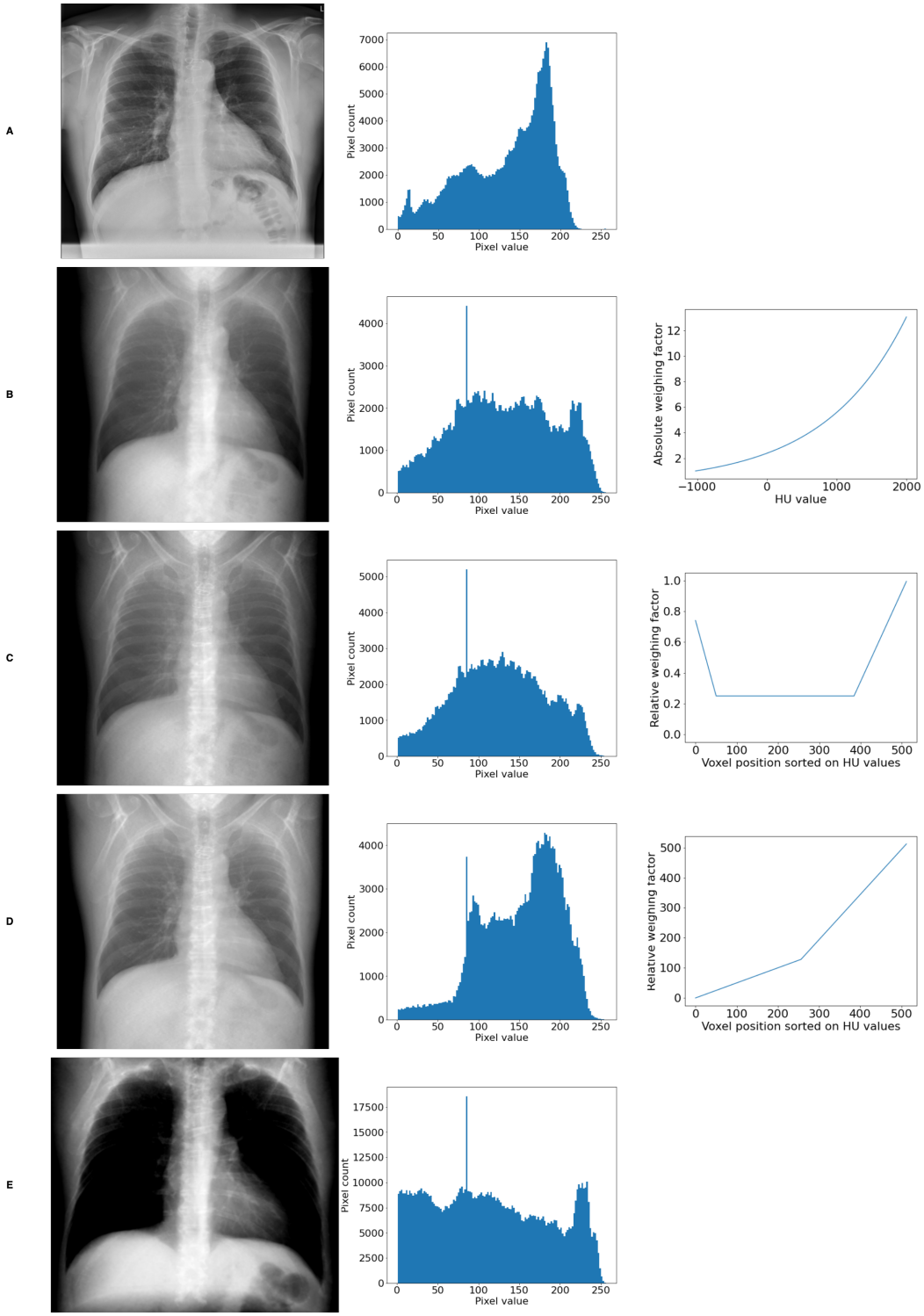


Fig. 12: Example DRRs constructed from an ULDCT patient case for which a CXR is also available. All images are evaluated at the same window-width and window-level settings as were present in the original CXR. Shown per column is the resulting image, the histogram for said image and the weighing factors used in the construction of the image if applicable. Shown per row is the original CXR (**A**), a DRR constructed using the approach by Campo et al. [33] (**B**), a DRR constructed using the approach by Carey et al. [35] (**C**), a DRR constructed using the approach by Meyer et al. [48] (**D**) and a DRR constructed using the approach by Unberath et al. [30] (**E**). The likeness of the DRRs to the original CXR is closely linked to the likeness in histograms, with row **D** showing the greatest similarities visually.

of the essence in intraprocedural settings. Abdellah et al. [49] express the importance of computing hundreds of DRRs in a matter of milliseconds to obtain the best possible image registration. In the work of Yoshino et al. [51], Yang et al [26] and Unberath et al. [30] such rapid computations suffice for the landmark detection of their respective applications. In this chapter this is not sufficient, and greater emphasis will be placed on the (diagnostic) image quality of the constructed DRRs.

The use of parallel based projection DRRs is more focused on the comparison between DRRs and existing X-Ray visualisations. In one example, Fuller et al. [34] looked into using DRRs from available CT data to assess the progression of flatfoot deformity compared to a regular X-Ray. Hamano et al. [52] compared plain hip radiographs to DRRs constructed from MRI data. Pyrros et al. [53] used DRRs and CT data to create new visualisations for the detection of lung nodules. What these works have in common is that a comparison is made between the diagnostic quality of a DRR and an existing conventional radiograph. By doing this, an emphasis is placed on the diagnostic quality of the constructed DRR, which is also a focus point of this chapter.

The primary difference between the applications discussed in the case of the point source and parallel based projections is the intended use of the DRR. In the case of point source based projections the DRR is a method to improve an already existing process. This is often an intraoperative or intraprocedural process which used DRR-type projections in the past, where DRRs are now able to mainly speed up procedures. In these cases the volumetric data is always obtained. For the majority of the parallel based projections the DRR is an explorative tool to speed up or aid in the interpretation of volumetric data. This too applies to certain point source based projection works, such as the work on the detection and classification of proximal femur fractures by Mutasa et al. [54]. It is unclear as to whether this resulted from the easy of implementation of a parallel-based projection compared to the point-source based projections or that another underlying reason was present.

IV. Methods

Several DRR construction methods have been identified that can be grouped into either a point source or a parallel based projection method. The goal is to identify which construction method is optimal and has suppport from clinical experts. To test this, an automated histogram-based evaluation method is proposed which is applied to the entire available dataset. Additionally, a clinical reader study is conducted in which radiologists are asked to fill out a questionnaire based on a selection of DRRs generated from normal cases. Both analyses are performed on known normal images, because the presence of (major) pathology can have a significant impact on a constructed DRR.

A. Implementation details

We identified four DRR construction methods from literature. These methods each describe a unique projection method. Overlap between these methods and other methods described in literature was ignored in favour of the paper which best described the projection method. Method 1 is a parallel DRR construction method based on the work by Campo et al. [33]. Method 2 is also a parallel DRR construction method based on the work by Carey et al. [35]. Method 3 is also a parallel DRR construction method bases on the work by Meyer et al. [48]. Method 4 is a point source DRR construction method based on the work by Unberath et al. [30].

We implemented each method in Python 3.8 according to the descriptions provided in the respective papers and with available online repositories if applicable, i.e. method 4 [30]. For computational efficiency the implementation for Method 4 made use of the pyCUDA Python package to enable GPU acceleration of DRR generation. The data is read from dicom files and is then stored as the compressed NIFTI file format for efficiency [55]. The DRRs, once created, as stored as a png image and are written back to a dicom file using the pydicom Python package. The SOPInstanceUID field was used to indicate differences in orientation.

B. Dataset

The work in this chapter is performed on a dataset that originates from the LUMC hospital. This dataset consists of 217 patient cases. Of these 217 cases 20 cases were deemed unusable due to the absence of images (n=8), the absence of radiological reports (n=6) or incomplete presence of images (either ULDCT or CXR was missing) (n=6). Patient consent had been waived by METC-Leiden Delt for the use of their data (number NL20210610001).

Every patient case consists of a FC08, or 'body', reconstructed ULDCT scan, a 'LUNG' or 'sharp' reconstructed ULDCT scan and two conventional CXRs; one lateral and one PA radiograph. For every patient case a radiological report was available for both the ULDCT as well as the CXR. Every case was examined for pathology, cross-referenced with the available report and subsequently sorted into one of two categories; no (active) pathology (n = 107) and (at least one type of) active pathology (n = 90).

C. Histogram-based evaluation

Direct image comparison between a DRR created from an (ULD)CT scan and a regular CXR is difficult. The primary reason for this is the difference in the manner in which both are obtained. A CT scan is taken lying down with arms stretched out behind the head. A CXR is taken standing up with arms around the detector at chest level. Because a CT is taken lying down fluids and air collections will show a different gravity sign compared to the CXR taken standing up. Furthermore, the resolution of a CXR is far greater than the CT scan.

A pixel-by-pixel comparison between a CXR and a DRR, even if they both originate from the same patient, is therefore unreliable. To provide some measure of how a DRR compares to a CXR the histogram of both images can be used. Because the CXR and DRR are globally aligned, a histogram comparison can provide a quantitative measure of image likeness.

Histogram comparison methods are generally either bin-to-bin or cross-bin comparisons [56]. The former is easier to calculate as pre-defined bins are compared to one another. In an 8-bit image this could for example be done with 256 bins. The latter is, however, more robust to variations such as lighting changes in images [57] but is more difficult to compute as the number of possible permutations between two sets of 256 bins is far higher. The key idea is to compare histograms in terms of overlap of their probabilistic distributions or by using a distance metric. Examples include metrics such as the Chi-Square metric [56], statistic correlation [**?**] or the Bhattacharyya distance [58]. We report means and standard deviations for each applied metric.

D. Clinical reader study

In order to obtain a clinically relevant assessment of the diagnostic quality of the four DRR construction methods, we conducted a clinical reader study with radiologists. The experts gave written informed consent with regards to their participation and the sharing of aggregate personal information. The clinical reader study was performed using the DICOM viewer MicroDicom[1] and was displayed on a 4K resolution screen. This is representative for the screens used in reading medical data. An example of the displayed DRRs is shown in figure 13.

In this study the participants were presented with six patient cases with known absence of pathology. The participants were informed of this. The six cases were presented to the participants in a randomised order. For every case the participant was shown the four constructed DRRs one by one in a randomised order. Before randomisation Method 1 is based on the work by Campo et al. [33], Method 2 on the work by Carey et al. [35], Method 3 on the work by Meyer et al. [48] and Method 4 on the work by Unberath et al. [30].

Participants were asked to provide a concrete rating on a 6-point Likert scale for a number of questions regarding the image quality and to provide a motivation for their choice in a text statement. These questions were set up to cover the relevant anatomical regions on an CXR. See table I for the original Dutch and translated English questions. Once all six cases had been shown, the cases were shown again in a newly randomised order where now the CXR corresponding to that case was included. Participants were then asked to judge which DRR best resembled the CXR.

[1]https://www.microdicom.com

Due to the limited number of participants in this clinical reader study, a focus was also placed on the qualitative feedback provided by the participants. The feedback was analysed in depth, and by using inductive category development as described by Mayring et al. [59] open issues were identified and used in future improvements [60].

TABLE I: Questions from the clinical reader study in original Dutch and translated English.

Dutch questions	Translated English questions
Ik kan deze DRR als een diagnostische thoraxfoto beoordelen	I can assess this DRR as a diagnostic CXR
Ik kan in deze DRR de weke delen diagnostisch beoordelen	I can assess the soft tissue in this DRR on a diagnostic level
Ik kan in deze DRR de ossale structuren diagnostisch beoordelen	I can assess the ossal structures in this DRR on a diagnostic level
Ik kan in deze DRR het mediastinum diagnostisch beoordelen	I can assess the mediastinum in this DRR on a diagnostic level
Ik kan in deze DRR de longen diagnostisch beoordelen	I can assess the lungs in this DRR on a diagnostic level

1) Participants: To assess the diagnostic quality of medical images, expert domain knowledge is required. Because of this, six medical professionals were recruited to participate in this clinical reader study. Of these six professionals, three are radiologists and three are residents in training with on average 7 (SD=5) years of experience reading CXRs in a medical setting. At the time of participation all professionals were employed by the LUMC hospital. The varied and unique backgrounds of the participants permits them to provide comments on the perceived diagnostic quality of the presented DRRs. At the same time, however, their expertise is sparse and their time is valuable, which makes them hard to recruit. A great emphasis is placed on their qualitative feedback to respect this.

V. Results

The results are presented for the histogram-based evaluation and the clinical reader study. The qualitative feedback obtained from the clinical reader study is discussed.

A. Histogram-based evaluation

The results for the histogram-based evaluation are shown in table II. A split is made between pathology, no pathology and both. Results did not differ significantly for any reported metric between these splits. Method 3 significantly scored best for the correlation metric on all splits. Method 2 significantly scored best for the Bhattacharyya distance metric on all splits.

B. Clinical reader study

The results from the clinical reader study are summarised in table III. Method 3, which is the 'softMip' approach by Meyer et al. [48] scored best on almost all categories. It was also picked as the DRR resembling the corresponding CXR most often, for 18 out of 36 comparisons. The senior participants tended to offer more detailed explanations and

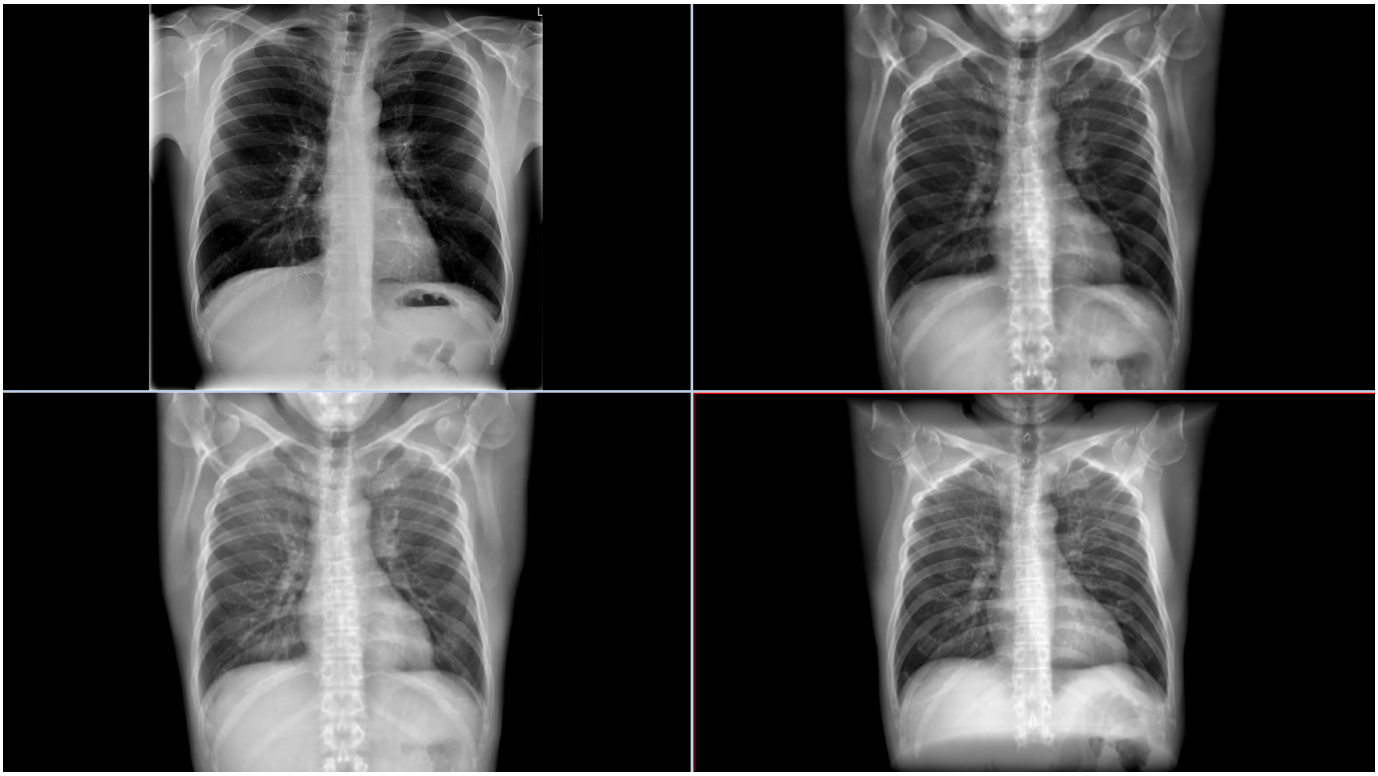


Fig. 13: A screen capture from the DICOM viewer MicroDICOM showing an example case with an original CXR in the top left, a DRR constructed with method 2 in the top right, a DRR constructed with method 3 in the bottom left and a DRR constructed with method 4 in the bottom right.

TABLE II: Results for the histogram based comparison. The analysis is performed on three splits of the dataset: one containing only pathology, one containing only no pathology and the entire dataset. Standard deviation is not reported for visual clarity. Shown are results for correlation and intersection (where higher is better) and Chi-Squared and Bhattacharrya distance (where lower is better). All metrics differed significantly compared to the original. In bold are methods where they both score highest (or lowest) on their histogram analysis and differ significantly (independent t-test, $\alpha = 0.05$) from other methods on the same test.

Grouping	**Histogram analysis**	**Original**	**Method 1**	**Method 2**	**Method 3**	**Method 4**
Pathology (n=90)	Correlation	1	$1.3 \cdot 10^{-1}$	$1.9 \cdot 10^{-1}$	$\mathbf{2.8 \cdot 10^{-1}}$	$8.0 \cdot 10^{-2}$
	Intersection	$1.0 \cdot 10^{7}$	$1.7 \cdot 10^{5}$	$1.8 \cdot 10^{5}$	$1.8 \cdot 10^{5}$	$1.0 \cdot 10^{5}$
	Chi-Squared	0	$1.1 \cdot 10^{7}$	$1.0 \cdot 10^{7}$	$1.0 \cdot 10^{7}$	$2.3 \cdot 10^{7}$
	Bhattacharyya distance	0	$5.9 \cdot 10^{-1}$	$\mathbf{5.8 \cdot 10^{-1}}$	$5.9 \cdot 10^{-1}$	$8.3 \cdot 10^{-1}$
No pathology (n=107)	Correlation	1	$1.5 \cdot 10^{-1}$	$2.0 \cdot 10^{-1}$	$\mathbf{3.2 \cdot 10^{-1}}$	$9.0 \cdot 10^{-2}$
	Intersection	$1.1 \cdot 10^{7}$	$1.8 \cdot 10^{5}$	$1.9 \cdot 10^{5}$	$1.9 \cdot 10^{5}$	$1.1 \cdot 10^{5}$
	Chi-Squared	0	$1.1 \cdot 10^{7}$	$1.0 \cdot 10^{7}$	$1.1 \cdot 10^{7}$	$2.4 \cdot 10^{7}$
	Bhattacharyya distance	0	$5.9 \cdot 10^{-1}$	$\mathbf{5.8 \cdot 10^{-1}}$	$5.9 \cdot 10^{-1}$	$8.0 \cdot 10^{-1}$
All images (n=197)	Correlation	1	$1.4 \cdot 10^{-1}$	$2.0 \cdot 10^{-1}$	$\mathbf{3.0 \cdot 10^{-1}}$	$8.0 \cdot 10^{-2}$
	Intersection	$1.0 \cdot 10^{7}$	$1.7 \cdot 10^{5}$	$1.8 \cdot 10^{5}$	$\mathbf{1.9 \cdot 10^{5}}$	$1.0 \cdot 10^{5}$
	Chi-Squared	0	$1.1 \cdot 10^{7}$	$1.0 \cdot 10^{7}$	$1.1 \cdot 10^{7}$	$2.2 \cdot 10^{7}$
	Bhattacharrya distance	0	$5.9 \cdot 10^{-1}$	$\mathbf{5.8 \cdot 10^{-1}}$	$5.9 \cdot 10^{-1}$	$8.1 \cdot 10^{-1}$

scored the DRRs less quickly. Moreover, the scoring by junior participants tended to vary more from image to image.

The written feedback provided by participants was analysed and grouped into overarching themes. The participating medical experts are referred to by E1, E2, ..., E6 and quotes are provided when comments are relevant to a certain theme. The number of individual experts who reference a certain theme or comment is denoted by n. The comments presented here from the experts have been translated from Dutch to English.

1) Resolution: All experts (n=6) made a reference to the resolution of the generated DRRs. E2 elaborated: *"The resolution of this DRR is more akin to a scout view from a CT scan than a CXR."*. E4 mentioned: *"I cannot be sure that I do not miss findings at this resolution."*. The resolution of the DRR was often mentioned in combination with the question on the assessment of the lungs in the DRR (n=4).

2) Noise: The level of noise present in the constructed DRRs was referenced by four experts (E1, E2, E3, E6). E1 stated: *"The level of noise in this DRR makes it difficult to assess the soft tissue."*. The level of noise played a role in answering the soft tissue (n=3) and the DRR as a diagnostic CXR (n=3) questions. In one case the level of noise was experienced as positive by E6: *"The increase noise in this image in combination with the increased lucency makes this image more pleasant to read"*.

3) Parallel compared to point source based projections: The underlying DRR construction method was not known to participants. Yet all experts (n=6) commented in some way on the difference between the parallel based and point source based DRRs. E3 mentioned: *"There seems to be a different distance between the dorsal ribs when comparing these DRRs."*. E4 continued: *"The chest wall seems to be further apart on this (i.e. point source) DRR compared to another (i.e. parallel based)."*. The perspective introduced by the point source based method was mentioned as a factor in answering the questions on the assessment of the DRR as a diagnostic CXR (n=3), the ossal structures (n=3) and the lungs (n=2).

4) Incomplete imaging and over projection: In one case the CT scan did not fully include the lung apexes. This was commented on by all experts (n=6). E1 elaborated: *"The scan did not fully include the apex of the lungs. I cannot know whether something was missed here."*. For this case, the fact that the apex was not fully scanned played a role in answering the DRR as a diagnostic CXR (n=2) and the lungs questions (n=6).

In the same case where Method 4 was used as a projection method experts (n=4) commented on the over projection of structures in the DRR. E2 noted: *"The over projection of the cranial end of the CT, in combination with the incomplete capture of the apex of the lungs makes this DRR impossible to read."*. The case in question with example DRRs is shown in figure 14. In two other cases experts (n=3) noted that the over projection of the cranial and caudal end of the scan was detrimental to the image quality.

5) Experimental setup: The clinical reader study was performed using the dicom viewer MicroDicom. This differs from the PACS viewer used in the clinical work setting. Multiple experts (n=4) noted the difference in dicom viewer influenced their decision making. E3 noted: *"I don't know if this image would look the same way in our own PACS."*. Image manipulation features such as zooming, or the setting of window-width or window-level were only used by one expert (n=1). The difference is keyboard shortcuts and window layout was also mentioned as playing a role (n=2).

The absence of pathology in the images was noted by all experts (n=6). E3 stated: *"Even though I can answer the questions for these cases knowing there's no pathology, the same answers will probably not apply when there is any pathology."*. E4 added: *"I'm curious what this will look like with certain pathologies."*. E3 continued: *"I'm the one responsible for not missing certain pathologies so I would have to see that it really works to be able to trust it."*. The trust in being able to see certain pathologies played an important role to the experts (n=4).

VI. Discussion

The goal of this chapter was to identify which synthetic CXR generation methods exist and how these were perceived by clinical experts. In this chapter we presented a review of related work and a novel evaluation of different DRR construction methods to answer this question. This evaluation consisted of an automated evaluation as well as an evaluation through a clinical reader study. In this section a discussion is presented on the obtained results, limitations with the evaluation are highlighted and subsequent steps are identified.

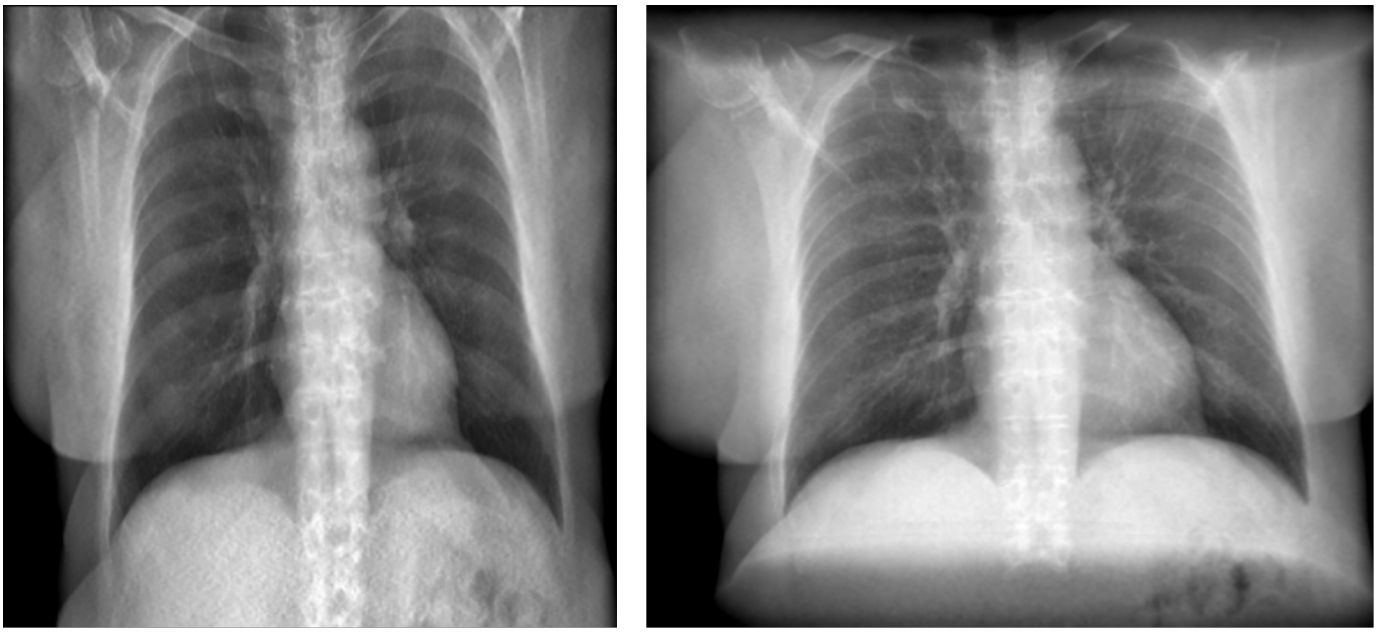
A. Histogram-based evaluation

To provide a quantitative and automated evaluation of the entire dataset that was available, a histogram-based evaluation has been performed. In this evaluation a comparison is made between the original CXR and the four methods with which the DRRs have been constructed. As can be seen in figure 12, the shape of the histogram can provide additional insight into how a DRR compares to a corresponding CXR. This type of evaluation is suitable as the DRR and CXR are not pixel-by-pixel comparable. This is due to the position of the patient when the ULDCT and CXR are taken respectively.

As shown in table II, method 3, the 'softMip' approach, scored the best overall on the correlation comparison metric. No significant effect was recorded for any other test save the Bhattacharyya distance metric on the 'all images' slice of the dataset. Visual inspection of both the 'softMip' DRRs as well as the histograms supports the suggested trend that the

TABLE III: Results from the clinical reader study. Reported are average scores with standard deviation from a 6-point Likert scale. In bold are the highest scores for every row. Method 1 is based on the work by Campo et al. [33], Method 2 on the work by Carey et al. [35], Method 3 on the work by Meyer et al. [48] and Method 4 on the work by Unberath et al. [30].

	Method 1	Method 2	Method 3	Method 4
DRR as a diagnostic CXR	3.0 [2.2 - 3.8]	3.4 [2.4 - 4.3]	**3.5 [2.6 - 4.4]**	3.1 [2.1 - 4.2]
Soft tissue on a DRR	4.0 [3.3 - 4.7]	4.3 [3.0 - 5.6]	**4.4 [3.1 - 5.6]**	4.2 [3.0 - 5.4]
Ossal structures on a DRR	3.3 [2.4 - 4.1]	**3.6 [2.5 - 4.7]**	3.4 [2.1 - 4.7]	3.3 [2.5 - 4.1]
Mediastinum on a DRR	3.6 [2.8 - 4.3]	**3.9 [3.0 - 4.8]**	**3.9 [3.0 - 4.8]**	3.7 [2.8 - 4.6]
Lungs on a DRR	3.1 [2.3 - 3.9]	3.0 [1.8 - 4.2]	**3.4 [2.6 - 4.2]**	3.3 [2.4 - 4.3]
Number of times identified as 'best'	5	9	**18**	4



A B

Fig. 14: Example case in which the lung apex was not fully imaged on the CT scan. **A**: example DRR projected by Method 2 where the apex of the lungs is not visible. **B**: example DRR projected by Method 4 where the apex of the lungs is not visible. Additionally, the boundary of the CT scan is over projected on the DRR at the cranial and caudal ends of the image.

'softMip' images best resemble the original CXR images.

This might not be entirely out of the blue, as the 'softMip' approach by Meyer et al. [48] specifically references its use in post processing of ULDCT projection data. The authors set out to combine the best of the edge sharpness of a Maximum Intensity Projection (MIP) and the image noise suppression of an Average Projection (AVG). In the presented analysis, a comparison is made with regular CXRs, for which the combination of suppressed image noise and sufficient edge sharpness is also relevant.

In related work histograms are primarily used in image enhancement [61], but they are also occasionally used in comparison or matching studies. Okada et al. [62] compared static and dynamic lung perfused blood volume images using histograms. Bottenus et al. [63] used histogram matching as a tool to normalise ultrasound images by referencing set standard images. This work is novel in that it compares histograms of the same patient case images across multiple modalities.

B. Clinical reader study

A clinical reader study was performed to assess the opinion of medical experts on the different DRR construction methods. As can be seen in table III this evaluation showed that method 3, the 'softMip' approach, scored the best across nearly all categories. This coincides with the result of the histogram-based evaluation. With the limited number of participants this therefore serves as a confirmation, thought not a statistical validation, of the results reported earlier.

In addition to scoring the questions on a questionnaire, the experts were asked to provide free-text comments to motivate their choices. The resulting feedback was compiled and sorted into five themes; resolution, noise, parallel

compared to point source based projections, incomplete imaging and over projection and experimental setup. The level of resolution in the DRRs was indicated to be a major complicating factor in reading DRRs as a diagnostic CXR. Resolution of constructed DRRs was also an issue for Mortani Barbosa et al. [32], who used super-resolution to bridge the gap between the resolution of the DRRs and their corresponding CXRs. This suggests that even though it is a serious issue, existing methods may be applied to resolve this.

The level of noise in DRRs was perceptually rated higher than in CXRs by our participants. The DRRs are reconstructed from ULDCT data and therefore contain more noise than a CXR. Yet the level of noise is also somewhat minimised by the method with which the DRR is constructed. As stated by Meyer et al. [48], the approach in method 3 has a de-noising effect by incorporating an AVG projection into the DRR construction method. The same applies to a lesser extent for the other construction methods where averaging over hundreds of voxels will have a de-noising effect. Despite this, the level of noise was noted as a detrimental factor in reading the DRR as a diagnostic CXR by four experts, two of whom commented this specifically for this DRR construction method. The level of noise also played a role in the work by Mortani Barbosa et al. [32], who applied an energy based normalization by Philipsen et al [64] to counter this. In addition to this, de-noising algorithms such as the work by Zhao et al. [65] could be applied to both the ULDCT data and the DRR in an attempt to resolve this.

The incomplete imaging and over projection, as shown in figure 14, is a major concern and was noted as such by multiple experts. This is caused by an external factor, i.e. an incomplete CT scan, but does provide valuable insight into the comparison between parallel and point source based DRR projection methods. The point source projection method introduces a divergence into the resulting DRR. The medical experts took note of this on several occassions and it influenced their decision making. To the best of our knowledge no comparable related work on this comparison exists.

The clinical reader study was set up using patient cases without pathology and using a DICOM viewer that differed from the regular PACS viewer in its interface and user interaction. Both factors were cited to influence the decision making of medical experts. In some cases it limited the interaction an expert would undergo with a certain image, given that they knew the outcome of the case or that they did not know how the software worked. The ramifications of not including pathology were This is a topic that will be tackled in a future chapter.

C. Limitations

The work presented in this chapter has a number of limitations. The automated histogram-based evaluation offers a limited quantitative analysis on the resemblance of DRRs compared to CXRs. Because these images are taken with patients in differing positions, a direct comparison is of limited use. Furthermore, only one of the proposed metrics showed a significant difference between the different DRR construction methods. This suggests that the difference between the different methods is marginal.

The clinical reader study was executed with a limited number of participants and a limited number of evaluated cases. Both were constrained by logistical limitations due to the sparse availability of medical experts. The validity of a future iteration of a clinical reader study would be improved by a greater number of participants.

The absence of pathology in the clinical reader study was stressed by participants as an important factor. The participants, operating from a position of responsibility for reading a patient case, were especially keen to know how certain projections look with regards to specific pathologies. It is possible that the preferred DRR projection method, the 'softMip' approach, does or does not display set pathologies very well. Future research will have to indicate this.

D. Future steps

The analysis of the clinical reader study has indicated several areas of improvement with regards to the construction of the DRRs. The resolution and level of noise of the DRRs were identified as major compromising aspects of the current approach. In related work super-resolution has already been applied to DRRs. In the field of de-noising there too has been research done with ULDCT data. This will be explored further in Chapter 3.

Additionally, the sparsity of medical expertise has highlighted the need for an automated evaluator for CXRs and DRRs. With such an 'automated radiologist', it would be feasible to rapid-fire alterations in the DRR construction methods or even optimise those with the automated response in mind. Chapter 2 will continue on this topic.

VII. Conclusion

In this chapter we identified and evaluated four DRR construction methods using both a quantitative histogram-based evaluation and a qualitative clinical reader study on patient cases without pathology. From both evaluations one DRR construction method emerged as 'best'. This approach, called 'softMip', showed a statistically significant difference in the automated approach and was preferred by the medical experts. The resolution and level of noise in the DRRs in addition to the presence of pathology were identified in the clinical reader study as primary areas of future work.

Using AI for disease classification in chest X-Rays and image evaluation in Digitally Reconstructed Radiographs

I. Introduction

The Chest Radiograph (CXR) is one of the most commonly performed radiological examinations in the world [66]. A CXR is quick and relatively cheap to produce and can be performed at limited radiation exposure [1]. The CXR has a reasonable sensitivity to a wide number of pathologies, which ensures it remains key in the diagnostic work-up of suspected chest pathologies [3]. In 2015, 142 CXRs were acquired for every 1,000 Dutch citizens (and 194 CXRs for every 1,000 EU citizens)[1]. This quantity of CXR examinations being performed has led to the accumulation of large datasets in the Picture Archiving Communication System (PACS) of various hospitals.

Deep Learning (DL) has become the go-to method for the (automated) analysis of medical images in the last few years [20, 67]. By showing hundreds of thousands of labelled images to DL models can be learned to perform a number of tasks with near human level performance. A 2021 review [68] divided these tasks into five key areas in which DL is applied to CXRs specifically. These include image-level predictions, segmentation, localisation, image generation and domain adaptation.

In image-level prediction, or image classification, DL models have achieved (near) radiologist level performance on a host of disease classes [66, 67, 69–71]. These developments have been made possible by the publication of several large accumulated CXR datasets. Since 2017 multiple repositories consisting of more than 100.000 labelled images each have been made publicly available [66, 70]. The images in these datasets are labelled primarily through Natural Language Processing techniques to parse disease classes from the concomitant radiological reports.

One key shortcoming in these successful approaches has been the lack of generalisation capability across different datasets. In numerous instances significant performance drops were noted when a model trained on one dataset was applied to another [72–74]. The field of domain adaptation attempts to shore up such differences through examples such as adversarial techniques [75] or joint training [76].

In the previous chapter we presented an analysis on the methods with which Digitally Reconstructed Radiograph (DRR)s can be constructed, and what clinical experts think of these. One of the key findings in this evaluation was related to the evaluation process itself. Presenting a representative selection of patient cases, with or without pathology, to medical experts was found to be very time consuming in chapter 2. Given the scarce availability of said expertise this quickly becomes unfeasible. Furthermore, such an evaluation would be a snapshot of a current iteration of images. Future insights into improvements of the DRR construction methods could mean that the evaluation has to be repeated.

These arguments greatly favour the introduction of an automated method of evaluating the DRR construction methods. Given the (near) peer performance of DL-models on the task of disease classification from CXRs, we propose to use these in evaluating the quality of different DRR construction methods. Such an evaluation would simultaneously provide an insight into the diagnostic quality of the different DRRs. We expect that tried and proven domain adaptation techniques such as joint learning can potentially even boost performance [76]. Because inference time with such models is negligible, it is possible to rapidly evaluate multiple DRR construction and processing techniques.

The following research question is therefore central to this chapter:

> *Can AI-models trained for chest X-Ray disease classification be used to evaluate the (diagnostic) image quality of Digitally Reconstructed Radiographs?*

This chapter is structured in the following way. In section II the background and related work on the use of AI-models in medical image classification. Section III describes how we applied and fine-tuned an established image classification approach to both CXRs and DRRs. This approach is evaluated and placed in context with literature in IV. A discussion on future work is presented in section V and a conclusion is provided in section VI.

II. Background & Related Work

In this background & related work section we provide an introduction to the topic of Deep Learning, it's application in medical image classification tasks and specific applications related to DRRs.

A. Deep Learning

Machine Learning (ML) is a specific programming technique that enables the extraction of data-driven rules from large

[1] https://vzinfo.nl/documenten/20210930datasterfteenverlorenlevensjaren2020oct

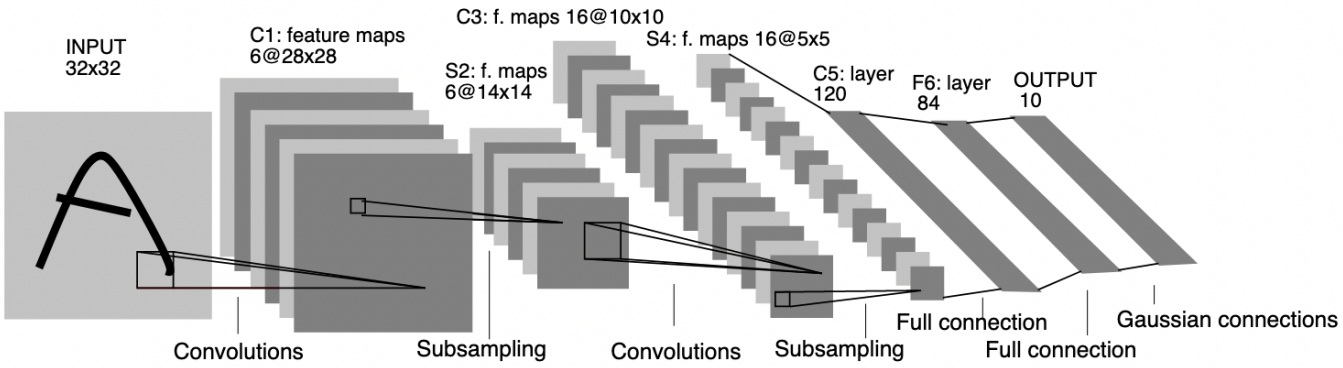


Fig. 15: Architecture of the LeNet-5 CNN designed for digit recognition. Every feature map represents a convolutional filter in a convolutional layer for which the weights are learned. Image used from [20].

numbers of examples without explicitly programming said rules [77]. For ML to work well adequate features describing the data have to extracted from these examples. This requires human domain expertise to craft feature extractors and apply them in a sensible manner. Deep Learning (DL) abstracts away from this approach by learning a representation from the raw data and subsequently crafting its own feature extractors in the form of a so-called neural network. These feature extractors are generally represented by successive layers which learn to compute an increasingly complex representation of the input data. By adding sufficient layers the network becomes 'deep' and is able to learn differentiation in complex data.

The learning aspect of DL models is driven by the ability of models to scale well to large datasets by making use of a tight integration of specialised soft- and hardware. This has enabled many DL models to beat the more traditional ML approaches. In most DL tasks the models are trained using a supervised learning approach. Here every sample of input data has a corresponding ground-truth label. In the medical context this can be a binary label stating the presence or absence of pneumonia on a CXR, but it can also be a complex segmentation of a cancerous lesion in a CT scan.

One of the most successful architectures of feature extractors is the Convolutional Neural Network (CNN). First used successfully by LeCun et al. [20], the CNN is able to encode spatial information with a degree of spatial invariance by making use of local receptive fields, by sharing or replicating model weights and spatial sub-sampling. Successive convolutional in the CNN layers encode increasingly complex representations of input data to the point where differentiation between input samples becomes possible. An example of the CNN architecture is shown in figure 15. As the understanding of CNN-based DL models grew so did the depth at which they were constructed. The intuition of 'the deeper the better' is in an ongoing struggle with issues such as vanishing gradients and hardware constraints. From the 19 layer VGGnet [78], the 50 layer ResNet [79], to the 201 layer DenseNet [80] advances are still being realised.

To benchmark these new model developments in object classification the ImageNet dataset is commonly used [81]. This dataset consists of over 1.2 million real-world images of 1.000 object classes. The race to be the best performer on this dataset has had a significant side-effect in kickstarting object classification in for example the medical imaging domain. Through a process called Transfer Learning, CNN-based models have shown to be able to transfer classification performance from one task to another [82]. By training a network on a large dataset such as the ImageNet dataset, the model is able to learn how to encode rudimentary visual information into a higher dimensionality representation that enables classification. In a separate second fine-tuning step this model is then further trained on a secondary, smaller, dataset consisting of the images related to the object classification task. Such models can then perform exceptionally well on smaller tasks without needing to train on large quantities of data.

In this chapter we apply the Transfer Learning principle to use the ImageNet pre-trained model weights from a DenseNet [80] model on our CXR classification task. The 'final layers' perform the classification of an image in one of the 1000 image classes present in the ImageNet dataset. Because we're dealing with only 14 disease classes, we cannot copy over those layers. When fine-tuning a model on both CXRs and DRRs on the same image classification task we can make use of the full architecture. A schematic overview of the application of Transfer Learning in this chapter is visible in figure 16.

1) Deep learning in medical related work: The deployment of the CNN architecture to medical tasks in combination with transfer learning from the ImageNet dataset has shown promising results in radiology [83, 84] ophthalmology [85], pathology [86] and dermatology [87]. Going beyond merely promising results, physician-level performance has been shown in the identification of diabetic retinopathy [88], breast

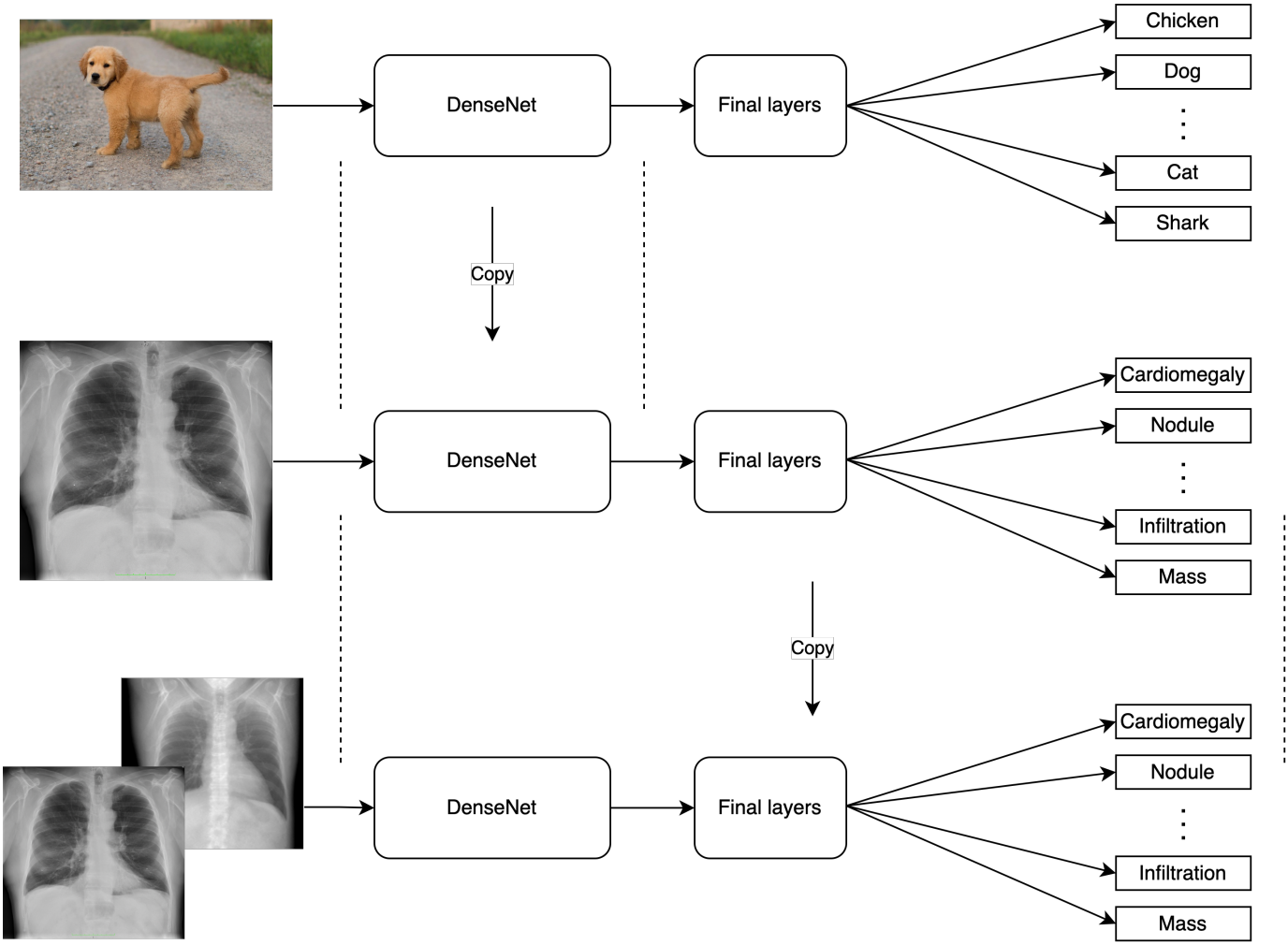


Fig. 16: A schematic overview of the model components that are copied over in Transfer Learning. The finaly layers of a model are not copied when a model is used for transfer learning on a different image classification task, as is visible in the second row. When the task is the same this does not apply, as is visible in the third row.

lesion detection [89] and spinal analysis using magnetic resonance imaging [90].

Despite such promising results, however, there are major remaining roadblocks barring clinical use of DL-based models in clinical practice. The chief obstacle in the development and deployment of AI is the availability of sufficient and qualitative data [91]. This was echoed by Tang et al. [92] who argue that data sharing and subsequently levelling the playing field between different sources are of great importance. Sourcing data from only one hospital may lead to models that generalise poorly to data from other institutions.

Another such obstacle is the context with which DL-based models generate predictions. In most circumstances this will be limited to an image or at best a set of images. In contrast, the clinical decision making happens using the clinical context consisting of patient examination, prior records and potential supplementary tests. Bridging this gap between the information that is available and the information that is used remains a challenge.

Finally, a growing importance is being placed on the so-called explainability of DL-based models. DL-based models are largely considered to be a 'black box', where it is difficult to retrace why a certain prediction or decision is reached. A number of different methods of explaining this decision making process have been proposed [93–95], where the focus is primarily on a visual support of a given prediction. It has been shown that this explainability is highly valued by experts in clinical practice [96]. Future developments should strive to keep this clinical desire in mind.

2) Deep learning in CXRs and DRRs: The ubiquitous availability of CXR data has provided the ideal circumstances for to the development of many DL models for disease classification [70, 83, 97, 98] on the CXR. Radiologist-level performance has been reported for a large number of different pathologies [98, 99], and radiology report generation has been shown to be feasible in a lab setting [100, 101].

DL models have also been applied to DRRs. Zhang et

al. [31] used DRRs computed from available CT data to train a Deep Learning (DL)-based model to segment covid infected regions in lungs on conventional CXRs. A similar approach was adopted by Mortani Barbosa et al. [32] where an additional image normalisation step was added. Campo et al. [33] used DRRs constructed from available CT data to train a DL-based model for the quantification of emphysema. Notably, Mortani Barbosa et al. [32] showed that their DL model using DRRs outperformed two human readers using CXRs.

As far as we know no direct comparison has been made with regards to a DL model applied to CXRs and DRRs taken and constructed for the same patient respectively. In this chapter we investigate this aspect further and show the effects of different image construction and processing methods on DL model performance.

III. Methods

As stated in the introduction, our goal in this chapter is to use an AI model to quantify the effects of applying image (post)-processing techniques to Digitally Reconstructed Radiographs. Before we can do this, we need an AI model that is capable of judging the image content in a meaningful manner. For this we make use of an image classification model as it learns to capture semantically relevant information in an image.

As we saw in section II there are a number of well-established approaches for the classification of disease in CXRs using DL. These approaches function well whenever they're trained on large and accurately labelled datasets. Our primary dataset of DRRs is simply too small to train an image classification model.

To compensate, we will validate one established approach on a larger, external dataset before fine-tuning and applying it to our own dataset. With this we can look into the effects of applying image post-processing techniques to DRRs on model performance. We start this section by discussing the different datasets we use in this and in subsequent chapters. We then present our approach on training, validating and applying this model.

A. Datasets

1) ChestX-ray 14, the NIH dataset: The work in this chapter is realised using a combination of several datasets. The LUMC dataset was discussed in detail in the previous chapter. The first of these is the ChestX-ray14, or NIH dataset [66]. The National Institutes of Health (NIH) dataset was originally released with 8 disease classification labels. This has since been expanded to 14 labels for each included PA CXR. The NIH dataset consists of 112.120 frontal PA CXRs including disease labels and is available for download in both PNG and DICOM formats[2].

The images in this dataset were collected in healthcare centres in the NIH network and represent the prevalence of disease in the patient community of these centres. A little over half of all included examinations are normal examinations. In the remainder the disease classes are not mutually exclusive where every labelled image has 1.6 ± 0.8 labels on average. A detailed overview of prevalence of disease classes is included in table IV.

The disease classes were originally extracted from radiological reports by Wang et al. [66] using natural language processing. They estimated these labels to be up to 90% accurate. Closer scrutiny of the labelling by Oakden-Rayner et al. [102] found that labels could be off in 10 to 30% of specific disease classes, putting into question the usefulness of the NIH dataset. In the disease classification work of Rajpurkar et al.[71] a new set of labels was made available. By training their model on a small subset of known high quality annotations they re-labelled a significant part of the NIH dataset. These labels have since been made available and cover the majority of the NIH dataset.

2) The CheXpert dataset: The CheXpert dataset [69] is in many ways a successor to the NIH dataset in that it is both bigger in numbers of images and more complex by adding uncertainty labels. This dataset consists of 224.316 labelled CXRs of 65.240 patients for an average of 2.3 ± 1.1 labels per image. The data was collected from the Stanford Hospital between 2002 and 2017 in both inpatient and outpatient centers. Almost all the data was made available publicly by the Stanford ML group[3] to host a competition on creating a model that can classify the Atelectasis, Infiltration, Pneumothorax, Consolidation and Edema disease classes. A private test set of 500 patients is used for this competition.

The CheXpert dataset was also labelled using NLP, but by allowing for the inclusion of uncertain classes the authors affirm that labelling can be more accurate than in the NIH dataset. Each image is assigned one of 14 labels, but these labels don't fully match the NIH dataset labels. A portion of them overlap, but uncertainty regarding the exact definitions of some of the disease classes means it is not possible to map 100% of all images between the NIH and CheXpert datasets. The labelling of the test set was done by an ensemble of 8 radiologists independently labelling the images where a majority vote was used to determine the final label. There is no public information available about the inter-rater variability.

3) LIDC/IDRI, lung nodule dataset: The Lung Image Database Consortium (LIDC) Image Database Resource Initiative (IDRI) is an initiative to share high quality annotated LDCT scans containing lung nodules. The LIDC IDRI dataset comprises 1.018 patients whose LDCT was annotated by four independent radiologists. Uniquely, this dataset contains a CXR for 297 of all the patients. This then enables a comparison between this dataset and the LUMC dataset where DRRs

[2]https://nihcc.app.box.com/v/ChestXray-NIHCC

[3]https://stanfordmlgroup.github.io/competitions/chexpert/

TABLE IV: Overview of the NIH, LIDC-IDRI, CheXpert and LUMC datasets. Nearly all data for the CheXpert dataset is publicly available.

Dataset	Number of images	Number of patients	Modality	Publicly available?	Annotation method
NIH	112.120	30.805	CXR	Yes	NLP parsing
LIDC-IDRI	297 / 1.018	1.018	CXR / LDCT	Yes	Committee experts
CheXpert	224.316	65.240	CXR	Yes*	NLP parsing
LUMC	197	197	CXR / ULDCT	No	Report and image parsing

TABLE V: Quantity of images available for the NIH dataset, the LIDC-IDRI dataset, the CheXpert dataset and the LUMC dataset split out over the disease classes specified in the NIH and CheXpert datasets where applicable. The disease classes are not mutually exclusive and a proportion of disease labels of the whole is included as %. The labels with a * are those which do not have a directly corresponding match between the NIH and CheXpert datasets.

Disease class	NIH	LIDC-IDRI	CheXpert	LUMC
Atelectasis	11.559 (10.3%)	-	33.376 (14.9%)	8 (4.1%)
Cardiomegaly	2.776 (2.5%)	-	27.000 (12.1%)	5 (2.5%)
Effusion	4.667 (4.2%)	-	-	12 (6.1%)
Infiltration	19.894 (17.7%)	-	-	2 (1.0%)
Mass	5.782 (5.2%)	-	-	8 (4.1%)
Nodule	6.331 (5.6%)	1.010 (100%)	-	23 (11.7%)
Pneumonia	1.431 (1.3%)	-	6.039 (2.7%)	0 (0%)
Pneumothorax	5.302 (4.7%)	-	19.448 (8.7%)	0 (0%)
Consolodation	4.667 (4.2%)	-	14.783 (6.6%)	15 (7.6%)
Edema	2.303 (2.1%)	-	52.246 (23.4%)	1 (0.5%)
Emphysema	2.516 (2.2%)	-	-	8 (4.1%)
Fibrosis	1.686 (1.5%)	-	-	1 (0.5%)
Pleural thickening	3.385 (3.0%)	-	-	14 (7.1%)
Hernia	227 (0.2%)	-	-	0 (0%)
No finding	63.016 (56.2%)	-	22.381 (10.0%)	111 (56.3%)
Enlarged cardiomediastinum*	-	-	10.798 (4.8%)	-
Lung opacity*	-	-	105.581 (47.3%)	-
Lung lesion*	-	-	9.186 (4.1%)	-
Pleural effusion*	-	-	86.187 (38.58%)	-
Pleural other*	-	-	3.523 (1.6%)	-
Fracture*	-	-	9.040 (4.1%)	-
Support devices*	-	-	116.001 (51.9%)	-

constructed from (ultra)LDCTs are compared to corresponding CXRs.

B. Image classification on CXRs and DRRs

As image classification model we implemented the ChexNet model architecture proposed by Rajpurkar et al. [70], now referred to as the '14-way model'. This architecture consists of a 121-layer DenseNet [80] which is pretrained on the ImageNet dataset [81]. The final fully connected layer was replaced with a layer corresponding to the number of disease classes in the NIH dataset. Similar to the original work a weighted cross entropy loss function [97] was used to train the network. The model was implemented using the open source Python framework Tensorflow[4].

The original labelling of the NIH dataset is not entirely reliable, as we discussed in subsection III-A. In this chapter we used a subset of 65% (or 72.787) images of the NIH dataset for which Rajpurkar et al. [71] provided updates labels. This subset largely represents a similar disease class prevalence distribution. Training, validation and test sets were constructed without patient overlap to prevent data leakage. Images were augmented using random cropping, rotation, channel shift but not a horizontal shift as this represents a semantically different medical examination result. All images were resized to a 224 by 224 pixel resolution and normalised using standard deviation and mean from the ImageNet dataset.

The model was trained using the Adam [103] optimiser with standard settings and a learning rate of 10^{-3}. Model training was done using a single GTX1080TI NVIDIA GPU where a learning rate decay callback with a patience of 5 epochs on the Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) of the validation set was used to determine model convergence. Five-fold cross validation was applied for model evaluation. Model training took 2.5 days.

1) Fine-tuning on a combined dataset of CXRs and DRRs: A secondary combined dataset was constructed from CXR images from the NIH datset and DRR images from the LUMC dataset to fine-tune the '14-way model'. Initially five CXR images were selected for every DRR where selection was stratified on disease class and patient ID. The resulting imbalance was alleviated using image rotation, image cropping and channel shifts on the DRR images. A second training

[4]https://www.tensorflow.org/

instance of the '14-way model' was started on this mixed dataset where starting model weights were taken from the trained '14-way model'. Here a weighted AUC score between the original CXRs, the DRRs and their combination was used to guide model convergence. All other model configurations remained the same.

C. Using the image classifier as image quality metric

In the previous chapter we saw that clinical experts perceive the various DRR construction methods differently with respect to the image quality. Given their scarce availability it was found to be unfeasible to repeat this evaluation for every alteration that could possibly be made. In this chapter we make use of the '14-way' image classification model as a metric of determining image quality instead. For this we evaluate several image quality adjustment methods on two separate datasets; the LUMC dataset and the LIDC-IDRI dataset. Based on the visual inspection of the histograms as in Chapter 1 we deem it sensible to try to either stretch the histogram contrast through a window width / window level preset or by re-sampling pixel values through some form of histogram equalisation. These methods are shown schematically in figure 17.

Window width and window level settings are one of the most commonly tuned parameters in the clinical practice of a radiologist. For CXRs specifically there are no set presets such as those that exist for CT. Additionally, the DRRs are stored as 8-big PNG images in contrast to the 12-bit CXR images, thus shifting potential preset values. We've opted to compare two presets based on the average shape of the DRR histograms; the moderate and aggressive windowing preset. The moderate windowing preset has a window level of 160 and a window width of 96 whereas the aggressive windowing preset uses a window level of 192 and a window width of 64. We also investigated a larger number of different window width and window level settings using a grid search approach.

Histogram equalisation has been applied as a medical image enhancement in numerous works [61, 104, 105] where subtle details can be enhanced by carefully applied tooling. For brevity two histogram equalisation methods are compared. A regular histogram equalisation method stretches the contrast in an image by assigning whatever lowest pixel value exists the value 0 and whatever highest pixel value exists the value 255. All pixels in between are re-sampled and additional contrast is added to the image. Contrast Limited Adaptive Histogram Equalisation (CLAHE) works on the

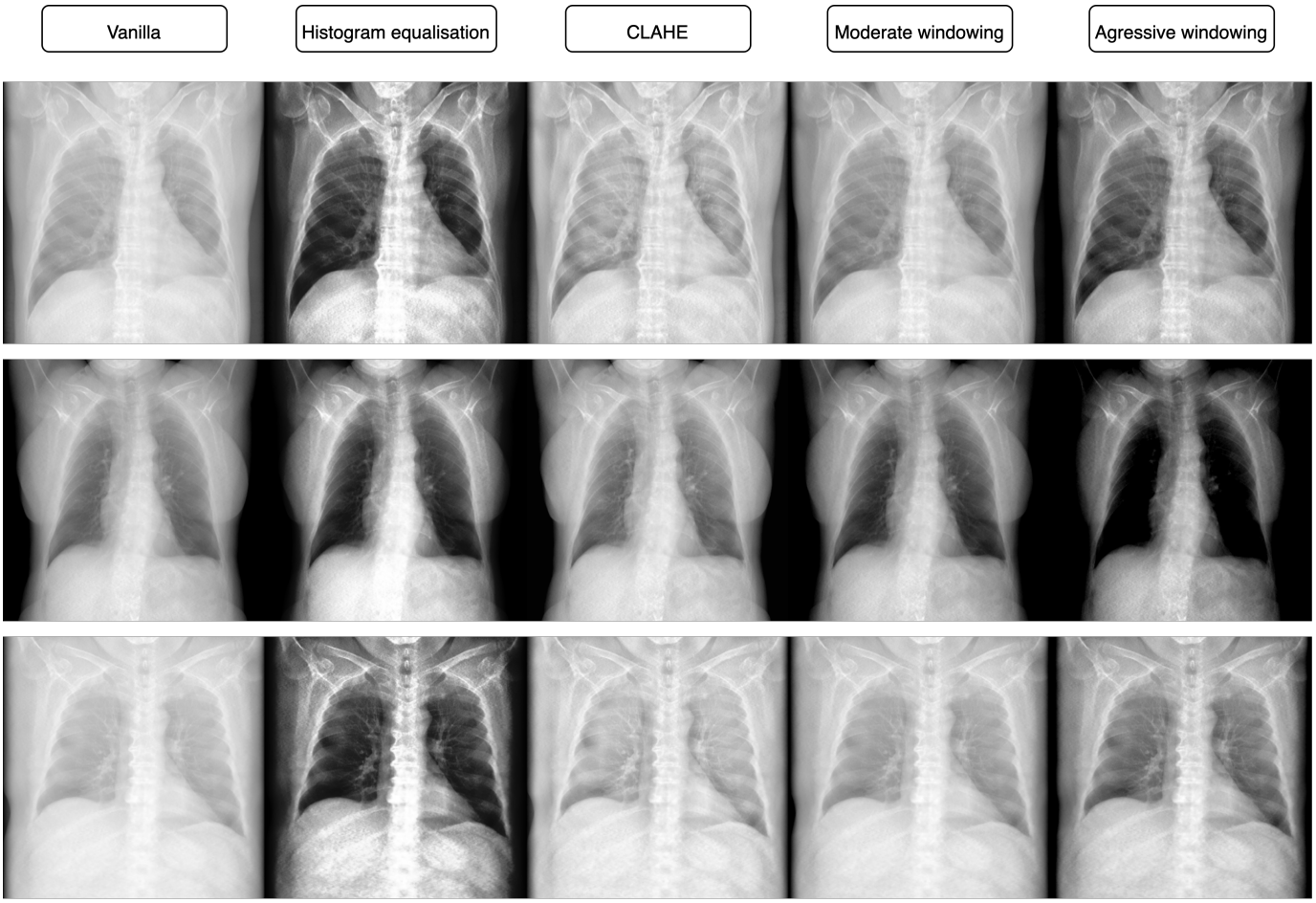


Fig. 17: Effects of the application of histogram equalisation, CLAHE, an aggressive window width / window level preset and a moderate window width / window level preset alongside a regular DRR for three cases.

same principal but only looks at a local neighbourhood for pixel re-sampling and applies a clipping limit to prevent artefact introduction [106].

IV. Results

A. Image classification on CXRs and DRRs

The '14-way' image classification model was evaluated on the unseen test set from the NIH dataset. The results for this are reported in table VI. The test set was also evaluated using the fine-tuned '14-way model' to evaluate the influence of further training the model using DRRs in addition to CXRs. As can be seen in the table the performance does not drop in a meaningful way across the different disease classes.

We also used a publicly available[5] version of the CheXpert model to evaluate our test set. This version only reports Atelectasis, Infiltration, Pneumothorax, Consolidation and Edema, so reported scores are limited to these disease classes. For comparison the original reported AUC scores from the CheXpert paper [70] are included as well. The authors also used the NIH dataset to train their model. Their reported scores outperform the other three model instances on all disease classes but Infiltration and Consolidation. For these classes Model 3 and Model 1 outperform the reported CheXpert scores respectively.

The performance of our image classification model is not as good as the original CheXpert model. We believe this may be down to a number of factors. Firstly, the test set on which the authors evaluated their model is not publicly available and consisted of a curated selection of 50 images per disease class. These were labelled by multiple radiologists and can therefore be considered as 'ground truth'. Given the noted objections in the data labelling quality of the original NIH dataset, it is possible that our reported performance is an underestimation of the actual performance. Additionally, the size of our test set (~ 2.400 images) was significantly bigger than the 420 images used in the CheXpert paper. These size differences could signify that our model gives a more realistic reflection of its performance.

Another factor that could affect model performance are model training times, which were limited by an overall AUC score. It is possible that another training regime could've led to increased model performance.

B. Using the image classifier as image quality metric

The trained '14-way' model and the fine-tuned version were applied to several instances of the CXRs and DRRs in the LUMC dataset. These instances were created using different image processing techniques as described in subsection III-C. The results for the evaluation of these instances using these two models are shown in table VII. Both Method 1 and Method 3 of the DRR generation methods outperform the original CXR in the vanilla visualisation, albeit by small margins. This suggests that these methods convey patient information equally well as a CXR.

The image processing techniques have varying effects. The application of image processing techniques to the original CXR always degrades model performance, which is not the case for the DRRs where occasionally marginal performance increases are seen. The grid search based alterations in window width and window level (G WW/WL) generally performs best as image processing technique. It is unclear whether these are definitive improvements or result simply from the unfamiliarity of the model with respect to that image processing technique. We note that over-fitting to the test set is a risk here.

[5]https://github.com/mlmed/torchxrayvision

TABLE VI: Evaluation results for the '14-way' classification model (**Model 1**), the fine-tuned '14-way' classification model (**Model 2**) and the CheXpert [70] model (**Model 3**) on the NIH test set. Only five categories are reported for the CheXpert model as the publicly available version only has these five outputs. The reported figures are AUC scores (with standard deviations) unless otherwise indicated. Included is a reference of the reported scores of the original CheXpert model (**Model 4**) on a different test set of the NIH dataset.

Disease class	**Set 1**	**Model 1**	**Model 2**	**Model 3** [70]	**Model 4**
Atelectasis	501	0.79 ($\pm$0.04)	0.79 ($\pm$0.03)	0.77 ($\pm$0.03)	0.81
Cardiomegaly	135	0.87 ($\pm$0.03)	0.85 ($\pm$0.03)	-	0.92
Effusion	199	0.85 ($\pm$0.02)	0.85 ($\pm$0.02)	-	0.86
Infiltration	883	0.72 ($\pm$0.02)	0.71 ($\pm$0.02)	0.79 ($\pm$0.02)	0.73
Mass	253	0.85 ($\pm$0.03)	0.85 ($\pm$0.02)	-	0.87
Nodule	289	0.77 ($\pm$0.02)	0.78 ($\pm$0.04)	-	0.78
Pneumonia	52	0.66 ($\pm$0.05)	0.66 ($\pm$0.05)	-	0.77
Pneumothorax	212	0.77 ($\pm$0.02)	0.71 ($\pm$0.03)	0.79 ($\pm$0.03)	0.89
Consolodation	241	0.80 ($\pm$0.02)	0.79 ($\pm$0.02)	0.82 ($\pm$0.02)	0.79
Edema	113	0.84 ($\pm$0.03)	0.82 ($\pm$0.02)	0.83 ($\pm$0.02)	0.89
Emphysema	103	0.91 ($\pm$0.02)	0.89 ($\pm$0.03)	-	0.94
Fibrosis	69	0.77 ($\pm$0.02)	0.73 ($\pm$0.02)	-	0.80
Pleural thickening	144	0.71 ($\pm$0.03)	0.71 ($\pm$0.02)	-	0.81
Hernia	13	0.70 ($\pm$0.06)	0.67 ($\pm$0.05)	-	0.92
Mean AUC	-	0.79 ($\pm$0.03)	0.77 ($\pm$0.03)	0.81 ($\pm$0.02)	0.84
Weighted mean AUC	-	0.78 ($\pm$0.02)	0.78 ($\pm$0.02)	0.79 ($\pm$0.02)	-

TABLE VII: Average AUC scores (and weighted average AUC scores) for the LUMC dataset and Nodule disease class accuracy scores for the LIDC-IDRI dataset. These scores are reported for the '14-way' image classification model (Model 1) and the fine-tuned '14-way' image classification model (Model 2). Shown in bold are the highest scores per row. The experiments are repeated for the vanilla DRR images, the DRR images where CLAHE was applied and finally the altering of the window width and window level settings on the DRR images (M WW/WL, A WW/WL and G WW/WL). Method 1 is based on the work by Campo et al. [33], Method 2 on the work by Carey et al. [35], Method 3 on the work by Meyer et al. [48] and Method 4 on the work by Unberath et al. [30].

Model	Dataset	Visualisation	CXR	**Method 1** [33]	**Method 2** [35]	**Method 3** [48]	**Method 4** [30]
Model 1	LUMC	Vanilla	0.80	0.81	0.75	**0.82**	0.80
		EQU	0.74	0.83	0.77	**0.84**	0.80
		CLAHE	0.75	0.78	0.67	0.76	**0.80**
		M WW/WL	0.75	0.83	0.77	**0.84**	0.78
		A WW/WL	0.66	**0.84**	0.77	**0.84**	0.68
		G WW/WL	0.77	0.85	0.77	**0.86**	0.79
	LIDC-IDRI	Vanilla	0.64	0.70	0.66	0.79	0.46
		EQU	0.78	0.85	0.87	**0.92**	0.53
		CLAHE	0.96	0.96	0.93	**0.97**	**0.97**
		M WW/WL	**0.78**	0.63	0.65	0.73	0.56
		A WW/WL	**0.96**	0.68	0.67	0.72	0.74
		G WW/WL	**0.90**	0.70	0.65	0.74	0.65
Model 2	LUMC	Vanilla	0.81	0.82	0.80	0.82	0.81
		EQU	0.77	**0.85**	0.83	0.84	0.81
		CLAHE	0.77	0.80	0.73	0.81	**0.83**
		M WW/WL	0.77	**0.83**	0.81	**0.83**	0.82
		A WW/WL	0.69	**0.85**	0.82	0.84	0.70
		G WW/WL	0.77	**0.86**	0.83	0.85	0.83

TABLE VIII: Detailed results for the classification performance on the different image post-processing techniques applied to the DRR constructed using method 3 [48]. Shown in bold are the highest AUC scores per row. The CXR results are reported as a baseline, with the different image post-processing techniques as a delta to this baseline. Both 'EQU', 'A WW/WL' and 'G WW/WL' have the highest net positive effect.

Disease class	Original CXR	DRR	EQU	CLAHE	M WW/WL	A WW/WL	G WW/WL
Atelectasis	0.87	-0.03	**+0.01**	-0.05	-0.02	-0.02	-0.01
Cardiomegaly	0.93	**+0.03**	+0.00	**+0.03**	**+0.03**	**+0.03**	**+0.03**
Effusion	**0.98**	-0.05	-0.05	-0.05	-0.05	-0.05	-0.04
Mass	**0.81**	-0.09	-0.01	-0.10	-0.08	-0.03	-0.02
Nodule	0.69	-0.03	**+0.01**	-0.02	-0.01	**+0.01**	**+0.01**
Consolidation	**0.83**	-0.05	**+0.00**	-0.09	-0.03	**+0.00**	**+0.00**
Emphysema	0.79	+0.08	+0.06	-0.03	+0.08	+0.04	**+0.09**
Pleural thickening	**0.72**	-0.05	-0.06	-0.09	-0.05	-0.03	-0.01
Mean AUC	0.81	+0.01	**+0.03**	+0.00	+0.02	**+0.03**	**+0.03**

The effect of fine-tuning (i.e. Model 2) generally results in an increase of model performance across all image processing techniques and all image types. Notably, this includes the original CXRs as well even though a different modality (the DRR) was mixed into the fine-tuning training set. This seems to suggest that the model has learned to generalise outside of its known image domain.

1) Evaluating on an external dataset: The performance of both model instances on the LIDC-IDRI dataset is reported in table VII. Here the accuracy is reported as every single image in this dataset has a known nodule. Therefore we cannot report AUC values. The accuracy is reported at the 95% specificity threshold determined using the '14-way' model and the NIH test set. The accuracy of the '14-way' model at this threshold was 45%. The results show that both histogram equalisation and CLAHE boost model performance compared to the vanilla version across all image types. This is in contrast with the LUMC dataset and may be explained by the fact that accuracy is reported instead of an AUC score. Furthermore, nodules are generally very small and the downsizing of input images to a resolution of 224x224 pixels may therefore lead to the loss of critical detail. The combination of down-sampling and image processing techniques mean these results are only an indication of the ability of this approach to generalise to external datasets.

2) Detailed evaluation of DRR performance: The detailed difference in model performance of the fine-tuned model on a number of different disease classes is shown in table VIII. We report CXR performance as baseline with detailed image processing performance as a difference. These DRRs were constructed using Method 3, or the 'softMip' approach. Reported here are only those disease classes for which at

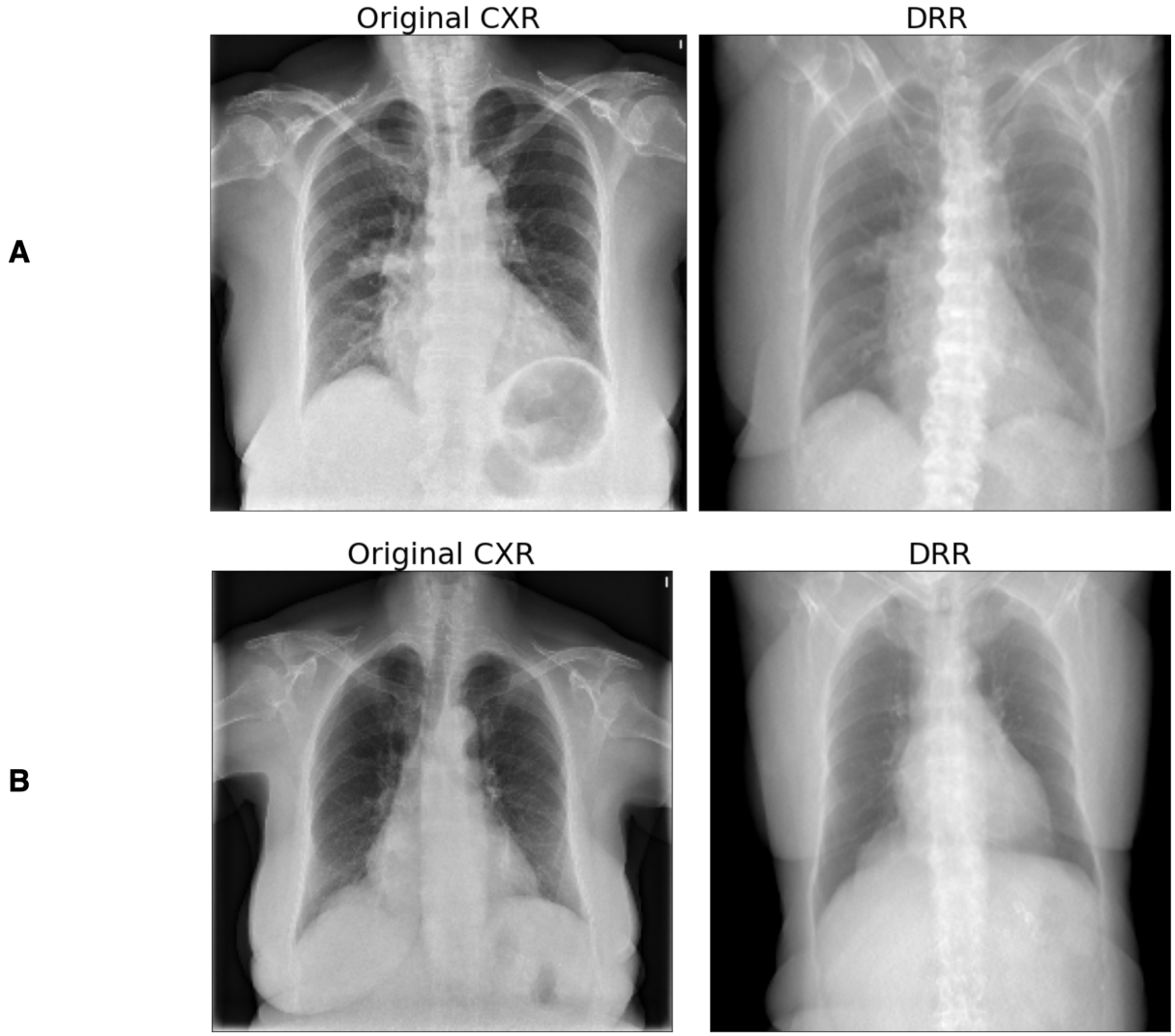


Fig. 18: **A** Patient case where the CXR was and the DRR was not recognised as containing a Mass. The Mass in this case was located at the right border of the heart. This presentation was mistakenly identified as Cardiomegaly in the DRR. **B** Patient case where the DRR was and the CXR was not recognised having Cardiomegaly. The enlarged heart is clearly visible in both images, although the CXR was mistakenly identified as containing a Mass.

least five cases were present in the test set. Here we also find that the grid search based alterations in window width and window level (G WW/WL) generally results in the biggest increase of model performance.

Across the different disease classes we find that the original CXR is always better recognised with regards to Cardiomegaly, and that the DRRs are always better recognised with regards to Mass. When investigating cases closely we find that labelling discrepancies in combination with overlapping findings might be the cause of this. In the first row (**A**) of figure 18 the CXR was and the DRR was not recognised as containing a Mass. Notably, the DRR was recognised as having Cardiomegaly. In the second row (**B**) of figure 18 only the DRR was recognised as having Cardiomegaly, but the CXR did trigger for Mass. Visually a case containing a Mass at the right border of the heart is hard to distinguish from Cardiomegaly at the resolution images are presented to a model (224x224).

V. Discussion

The goal of this chapter was to identify whether AI models could be used to evaluate DRRs and the effect of image post processing techniques. We've shown that it is possible to apply an AI model trained on CXR images to a dataset of DRR images and we've investigated the effects of several image processing techniques on the model performance. Furthermore, we've shown that fine-tuning an image classification model with a mixture of CXRs and DRRs improves the performance on both CXRs and DRRs.

A. Image classification performance

The '14-way' disease classification performance is considerably worse than the reported classification

performance of the CheXNet model by Rajpurkar et al [70]. At the same time, the performance of this model we obtained using a publicly available version of this model[6] was also worse than the reported classification performance. In our evaluation the size of the test set was equal, but in our approach we used cross validation. We do not, however, believe that this explains the difference in performance.

In training our model, we followed the same set-up as was described in the original paper but could not match their performance. We believe this may primarily be down to the training time. In our approach the model convergence was determined using the AUC as callback. This was not explicitly described in the original paper, where a different convergence metric could've led to a different model performance.

We showed that DRRs constructed using the 'softMip' approach are at least as good as the original CXRs with regards to disease classification model performance in both the LUMC and LIDC-IDRI datasets. Mortani Barbosa et al. [32] found DRRs to be as good as CXRs with regards to COVID-19 disease classification. Here we show that this extends to other disease classes.

B. Validating on an external dataset

We validated our disease classification model on the external LIDC-IDRI dataset. We chose to do this as DL performance is known to fall off on different datasets than what a model was trained on [68]. Though it is of limited size and only labelled for one class, the LIDC-IDRI dataset helped us show the ability of this approach to generalise. Our reported performance is in line with the review performed by Li et al. [107] on different related works using this dataset. We found the performance to be highly susceptible to changes in window width and window level settings. We believe this may be down to the generally higher than normal HU values of nodules. As the window width and window level settings change, their presence may become binary in that they are either visible well or not at all.

C. Fine-tuning on DRRs and CXRs

In our fine-tuning approach we showed that it is possible to boost the performance of a classification model trained on CXRs when applied to DRRs by continuing to train it on a combined dataset of the two. At the same time the model performance did not drop notably on the original CXRs. A similar positive effect was seen by Mortani Barbosa et al. [32], who achieved the best performance on a COVID-19 CXR classification problem using an ensemble of CXRs and DRRs. We showed that this can be extended to multiple disease classes.

A closer inspection of both the Mass and Cardiomegaly disease classes revealed a number of cases where a mass located close to the heart caused confusion in the model performance. We speculate that such an overlap could also be possible in general for the Infiltration, Effusion and Consolidation disease classes. These disease classes have very similar clinical presentations and are generally hard to distinguish even for experts [102].

D. Limitations

There are several limitations related to the datasets that were used in this chapter. The NIH dataset is known to contain significant labelling errors. The re-labelling efforts will alleviate these issues somewhat, but without clear, publicly available definitions of what constitutes a disease class and how this maps to the extraction process from a radiological report significant errors will most likely persist. Any model performance is therefore likely to in part be a reflection of the ability of a model to learn erroneous labels and not necessarily a reflection of real performance.

The LUMC dataset was labelled using an approach that tried to follow the described approach in the NIH dataset. Because this was not done by a board certified radiologist, it is not unlikely that labelling errors were introduced in this dataset as well. This would compound on existing errors and diminish reported AUC scores on this dataset.

Finally, the availability of datasets where both an (ultra)LDCT and a CXR are available for the same patient is extremely scarce. In fact only the LIDC-IDRI and LUMC datasets contain this unique combination. The LIDC-IDRI dataset was only labelled for nodules and therefore no comparison could be made for any of the other disease classes. Constructing a larger dataset containing this combination of modalities could greatly benefit the understanding of the applicability of image classification models across these closely linked imaging domains.

VI. Conclusion

In this chapter we've presented an automated image evaluation pipeline by using a CXR-trained image classification model as evaluator. We demonstrated that this model generalises well to DRRs and that fine-tuning does not degrade model performance on either imaging domain. We used our disease classification model to evaluate image post processing techniques to find that both histogram equalisation and alterations in window width and window level settings can boost disease classification performance.

[6]https://stanfordmlgroup.github.io/competitions/chexpert/

Using AI to generate realistic chest X-Rays and transform DRRs

I. Introduction

The widespread availability of Chest Radiograph datasets has contributed to the development of AI models that can achieve near radiologist performance on a host of different pathologies [66, 67, 69–71]. Despite these successes, however, concerns have been raised regarding the drop in performance when a model trained on one dataset is applied to another dataset [68, 72–74]. One way to resolve such issues is to construct even larger, multi-centre datasets. Such an approach is very expensive and is fraught with issues like patient privacy, labelling standards and storage responsibilities [108, 109].

A different approach that has already shown successful applications of AI models is to use generated images. The use of the Generative Adversarial Network (GAN) architecture has enabled the creation of realistic human faces at high resolution levels [110, 111]. These developments have also been applied to medical imaging domains with examples such as the generation of realistic skin patches [112] and MRI to CT synthesis [113]. The use of GANs has the potential to alleviate issues like class imbalances in current datasets without having to worry about privacy concerns regarding patient data [108, 109, 114–116].

In the previous chapter we found that it is possible to analyse DRRs using DL-based models trained on CXRs. This helped us to evaluate the effects of varying window width and window level presets in addition to different image post-processing techniques. Despite this, we noted that this did not necessarily lead to an 'optimal' visualisation in the context of an AI model.

In this chapter we propose to approach such an 'optimal' visualisation by using a Generative Adversarial Network (GAN). These network architectures have successfully been applied to the generation of realistic (medical) images. Such networks function by mapping random latent variables to realistic images in the target domain. Once this behaviour is learned, we propose to reverse the mapping process and find the optimal latent variable representation of an image. Specifically a DRR image. This would enable us to 'create' a CXR for a given DRR.

The following research question is therefore central to this chapter:

> *Can models be used to generate realistic CXR and can they subsequently facilitate the generation of a CXR visualisation for a DRR?*

This chapter is structured in the following way. In section II the background and related work on generating images using GANs is discussed. Section III describes how we trained our image generation model and which evaluations are relevant. The results for these evaluations are reported and placed in context with literature in IV. A discussion on future work is presented in section V and a conclusion is provided in section VI.

II. Background & Related work

A. Generative Adversarial Networks

The Generative Adversarial Network (GAN) architecture was initially proposed by Goodfellow et al. [117] and consist of two competing networks. These networks are the discriminator D and the generator G which both work on some dataset X. The task of the discriminator is to be able to discriminate between real images: $x \in X$ and fake images: $\hat{x} \in \hat{X}$ such that $D(x) = 1$ and $D(\hat{x}) = 0$. At the same time the generator attempts to find a mapping from some latent variables $z \sim p_z(z)$ to generated fake images $\hat{x} \in \hat{X}$ such that the discriminator is unable to tel the difference between real and generated data. A high level overview of the generator and discriminator architectures is included in figures 19 and 20 respectively. The relationship between generator and discriminator is expressed in the following value function V:

$$\min_G \max_D V(D,G) = \\ E_{x\sim p_{data}(x)}[log(D(x))] + E_{z\sim p_z(z)}[1 - log(D(G(z)))] \tag{8}$$

To train a GAN this value function is optimised in a careful balancing act between the two adversarial networks. The discriminator is provided known real images to enable the computation of a gradient for both the discriminator and the generator. One common failure point in this process is the so-called mode collapse in which the generator defaults to a single mode of image generation. The training process is shown schematically in figure 21.

To effectively train GANs enormous quantities of data are required. This helps show both the discriminator and generator what variation exists and what variation is acceptable. As a matter of reference the current state-of-the-art GAN architectures are trained using more than 100.000 high quality images.

1) GANs in related work: In the medical context GANs have seen use in applications such as MRI reconstruction [118], CT denoising [119], CT to MRI synthesis [113], image segmentation [120] and (un)conditional synthesis [112]. In

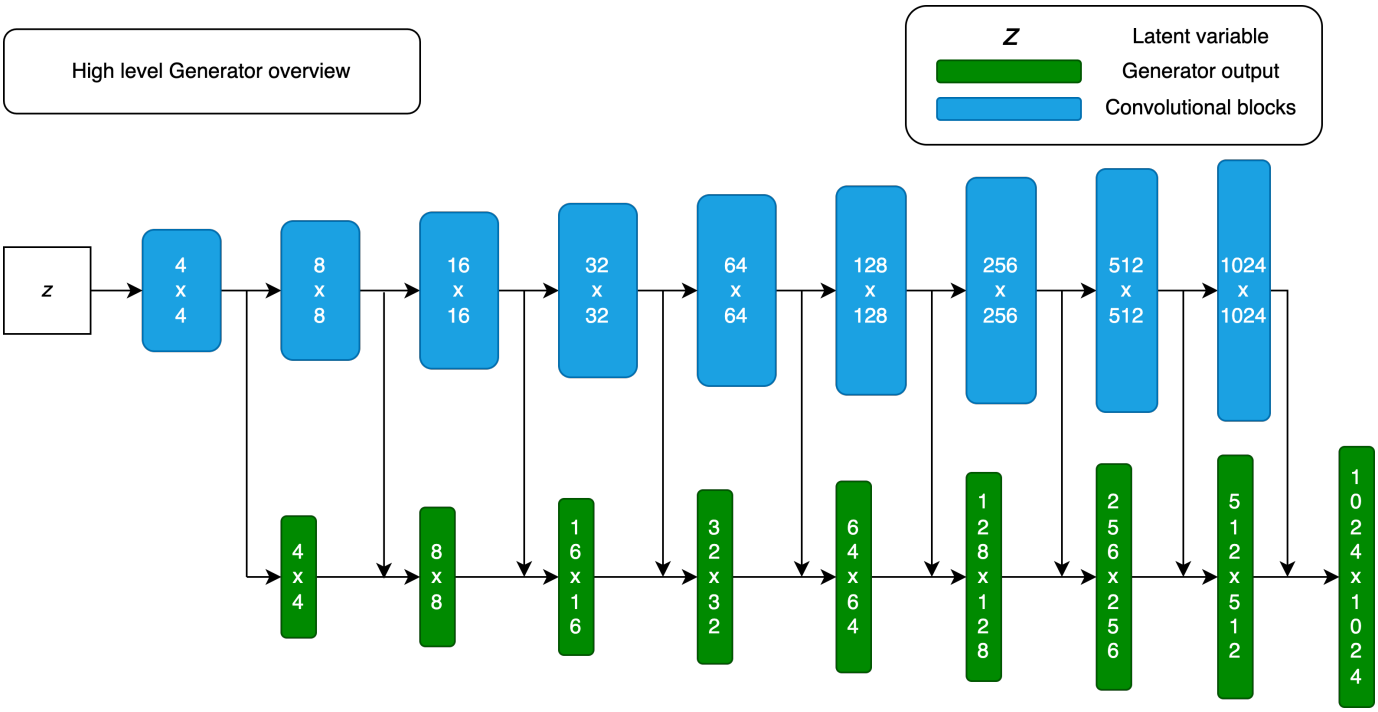


Fig. 19: High level overview of the generator of the PGGAN architecture. Images are initiated with a random sample z from the latent variable space. Successive convolutional blocks reshape and up-sample the generated image by combining prior resolutions and the output of the corresponding convolutional blocks. The width of the convolutional blocks is a visual indicator of the number of convolutional layers used, i.e. the higher resolution blocks use fewer layers than their lower resolution counterparts. The points at which the arrow connections meet represent a concatenation of layers.

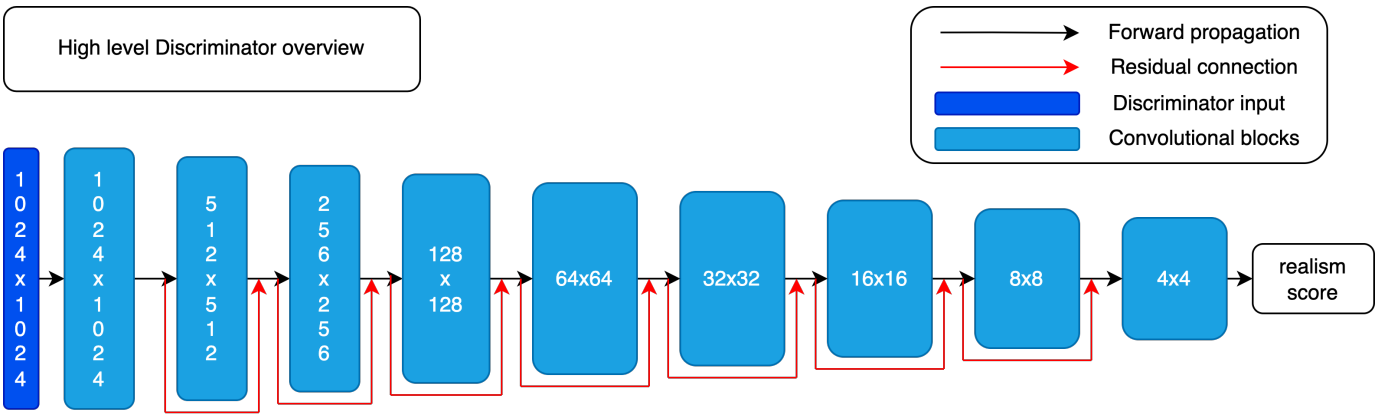


Fig. 20: High level overview of the discriminator of the PGGAN architecture which computes a realism score for a given input image. The discriminator uses an image of the highest output resolution by the generator as its input. Successive layers of convolutional blocks reshape and down-sample the image where the lower resolution layers use more convolutional layers per block than the initial higher resolution layers. Shown in red are the characteristic residual connections of the ResNet [79] architecture which 'skip' the convolutional blocks. The points at which the arrow connections meet represent a concatenation of layers.

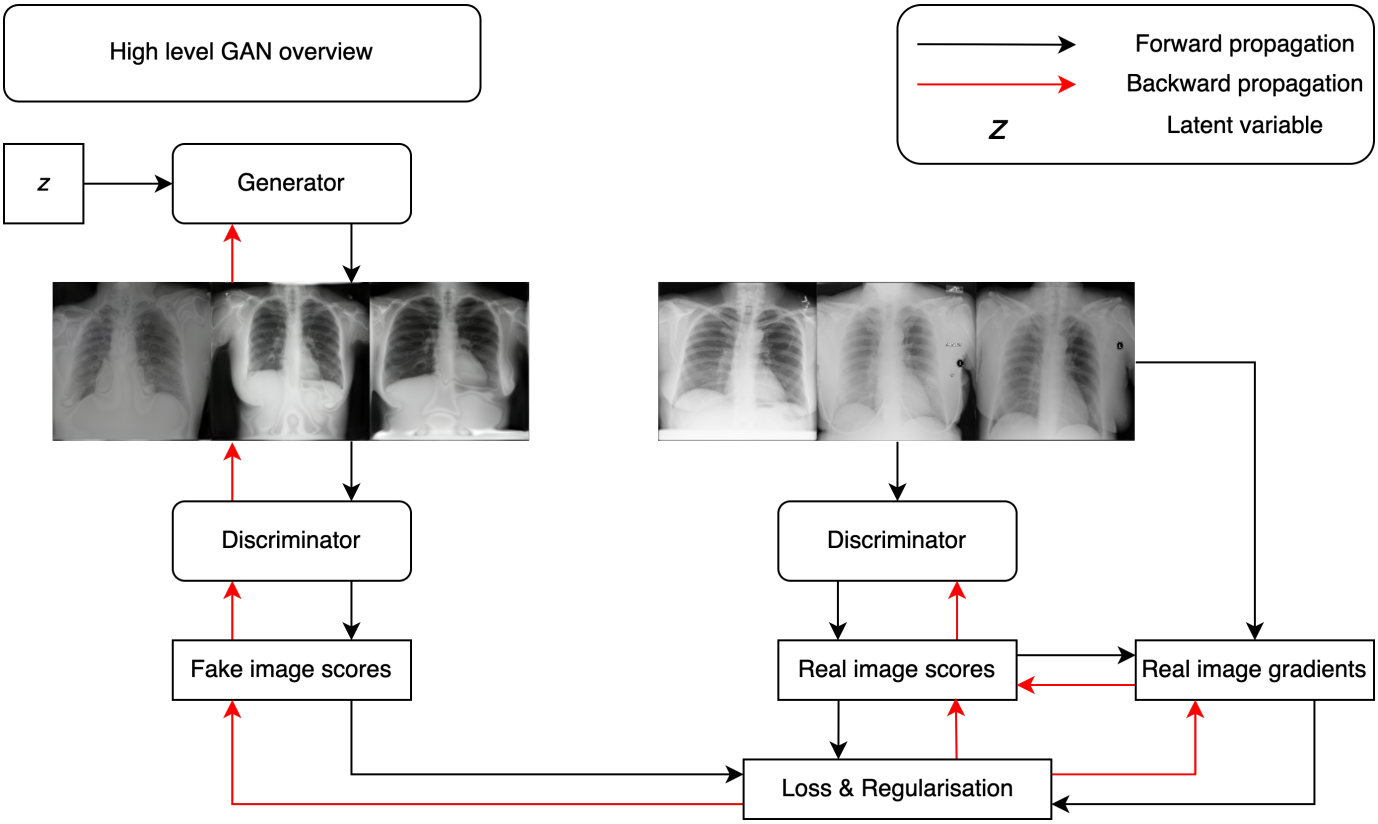


Fig. 21: Overview of the training process of a GAN architecture. Fake images (on the left) are generated from a latent variable representation z by the generator. Real images (on the right) are sourced from a dataset to provide a comparison between generated and real images for the discriminator. This comparison is used to compute a loss function which can then update both the generator and discriminator through backward propagation of the loss function.

most of these applications GANs are employed to alleviate the problem of data scarcity and dataset imbalances by providing synthesised examples of whatever data is missing [121].

To handle such increasingly complex tasks additions to the original fully-connected GAN architecture have been proposed. Convolutional layers were added by Radford et al. [122] to improve GAN resolution. Further progress was made by using progressively growing generated sizes [123], style transfer techniques [110] and alternate loss functions [124].

In this chapter we employ the Progressively Growing GAN (PGGAN) architecture proposed by Karras et al. [123]. In this architecture both generator and discriminator grow progressively in their respective resolution sizes. During training this allows to model to increasingly focus on finer details in the generated images and simultaneously speeds up the initial part of training. The incorporation of minibatch discrimination effectively counters the possibility of a mode collapse by introducing the variability per minibatch as a training variable to the generator.

Further improvements to the PGGAN architecture have been proposed in the StyleGAN [111] and StyleGAN2 [110] architectures. By adding stochastic noise to the generator further variation is introduced besides the initial latent space sample. Furthermore, these architectures were shown to be less prone to artefact generation at higher resolutions. Despite these advances, these architectures are computationally unfeasible to employ.

2) GANs in synthesing CXRs: The application of a GAN to a specific (medical) imaging domain has the potential to bring a host of benefits. Problems inhibiting the training of DL models due to data scarcity or class imbalance can be tackled by generating missing data [121]. This data can be generated and subsequently shared without patient privacy concerns regarding data sharing.

GANs have been applied as data augmentors in the training of a CXR classification model by Salehinejad et al. [116]. They reported improved test set accuracies by including generated CXRs in their training set. These results were reproduced by Sundaram et al. [125] and Moradi et al. [126] who reported similar effects on differing datasets. The authors agreed that class balancing played a large role in the improved classification performance.

Efforts to directly generate DRR images are not likely to succeed. The limited availability of CT data, compared to CXR data, means that only a very sparse variety of

DRRs could be generated. It is, however, possible to use the intuitions behind the StyleGAN architecture to make DRRs more CXR-like. To the best of our knowledge this has not yet been researched. In this chapter we investigate this possibility, where we have to invert the generator to obtain a latent mapping of our input image such that we may make it more CXR-like.

III. Methods

The goal of this chapter is to investigate whether it is possible to create an 'optimal' CXR-like visualisation of a DRR. To achieve this we need to take two successive steps. Firstly, we need to create a model that is able to generate realistic CXRs from a random latent variable z. As we saw in section II there are a number of established approaches using GANs that can achieve this.

Secondly, we need to be able to find the latent variable representation of a given DRR image. If we can achieve this, we can obtain the generated CXR that best resembles the DRR. This depends on the CXR-generation model being able to represent normal varieties in sex, anatomy and pathology as they occur in the DRR dataset.

A. Synthesising CXRs

We implemented an upgraded PGGAN model in Tensorflow based on the open-source code provided by Karras et al. [123][1]. Insights from the StyleGAN [111] and StyleGAN2 [110] were incorporated into the PGGAN model architecture. These include skip-connections in the generator instead of solely progressive growing, equalised learning rates, the addition of stochastic scaled noise to each channel and a reduction in perceptual path length to improve image quality.

The NIH dataset was converted into TFRecords and hosted in a Google Cloud storage instance for data access. Model training was initialised with random weights starting at a 4x4 resolution and 512 convolutional layers. Layers were kept constant as resolution doubled through 64x64 resolution, after which convolutional layers were halved to reach 32 at 1024x1024 resolution. We utilised the softplus loss as the WGAN-GP loss has been shown to not converge at higher resolutions [124] and this drives the generator to improve on its worst samples. Additionally, pixel normalisation, minibatch discrimination and lazy regularisation (once every 8 steps) were applied. Latent variables were generated from a normal distribution $z \sim N(0, 1)$.

The model was trained using a distributed training strategy on hardware available through a Google Colab instance consisting of 8 TPUs. We trained the model by showing 2.048 images per epoch to each distributed training instance until 26 million images were shown. Model training was visually inspected using a custom visualisation callback which smoothed over visualisation weights by applying weight decay. Total model training took 9 days.

[1]https://github.com/tkarras/progressive_growing_of_gans

1) Creating 'optimal' CXR images: The combination of a CXR image generation network and a CXR image classification network enables the creation of 'optimal' CXR images for each specific class. To achieve this we 'reversed' the generator of the PGGAN network and used the '14-way' model ($D*$) instead of the regular discriminator. For a given latent variable input z a classification output is produced for a specific class:

$$D*(z_{cardiomegaly}) = c_{cardiomegaly} \tag{9}$$

A loss function L can then be defined with respect to a target value T:

$$L_{cardiomegaly} = \|T_{cardiomegaly} - c_{cardiomegaly}\|^2 \tag{10}$$

Where the original input z can now be updated using gradient descent by backward propagation of this loss through the fixed discriminator and generator networks:

$$z_{cardiomegaly} = z_{cardiomegaly} - \Delta_z L_{cardiomegaly} \tag{11}$$

This process is shown schematically in figure 22.

2) Obtaining latent space representations of DRRs: The process we described for generating an 'optimal' image with respect to a disease class can be taken a step further. Instead of directing the generator towards a disease class we used a discriminator to calculate the similarity between the generated image and a target DRR image.

In this case we used a pre-trained VGG16-network to extract features from both the generated and the target image. These features were computed by extracting the penultimate layer of this model. Using both the generated image ($z*$) representation $D_{vgg}(z*)$ and the DRR image representation $D_{vgg}(drr)$ we computed a loss function to optimise the latent variable:

$$L_{drr} = \|D_{vgg}(drr) - D_{vgg}(z*)\|^2 \tag{12}$$

Where we updated the original latent variable $z*$ using gradient descent by backward propagation of this loss through the fixed discriminator and generator networks:

$$z* = z* - \Delta_z L_{drr} \tag{13}$$

To compute the VGG16-features we down-sampled both the generated and DRR images to a 224x224 size using bicubic interpolation [127].

B. Evaluating image classification performance using GAN-train and GAN-test

One of the methods with which the image quality of generated images can be assessed is by using the GAN-train and GAN-test principles proposed by Shmelkov et al.[128]. In the GAN-train approach an image classifier is trained solely on generated images and then tested on some labelled test set of real images. The GAN-test approach flips this by evaluating an image classifier on a test set of generated images. We implement these approach by using the trained '14-way' model classifier described in the previous chapter to label a train

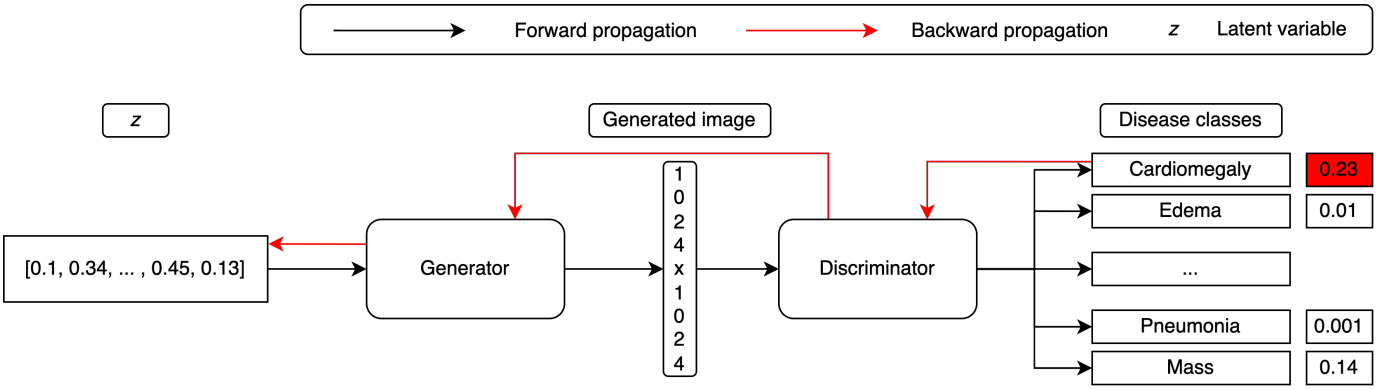


Fig. 22: Overview of the process of optimising a latent variable sample z with respect to some disease class c. This is achieved through backward propagation of the loss with respect to a class c through **fixed** discriminator and generator networks. This cyclical process is repeated until the visual result is adequate or some set threshold with respect to the disease class c is reached.

and test dataset of generated CXR images that were obtained following the approach described in subsection III-A. We use the applied cross validation to create an ensemble voting committee to provide labels to the generated training images where a majority vote of 3 suffices to assign a label. An image classifier is then trained similarly to the methods described in the previous chapter.

IV. Results

In this section we report our results on the generation of realistic CXRs. In certain subsections we make use of the '14-way' classification model. This is the disease classification model as described in the previous chapter.

A. Synthesising CXRs

The results of a sample of randomly generated CXR images by the trained PGGAN generator is shown in figure 23. The general characteristics of the NIH dataset are represented in the generated images. Both male and female CXRs are generated at varying levels of image exposure. Empirically, the majority of the generated images contain all major thoracic organs. In most images the common anatomical position of these organs can be seen, with landmarks such as the heart shadow, the stomach bubble, the slightly higher diapgrahm for the liver and the aortic knob to name a few. These structures occur with a normal variety in anatomy. The ribcage and other ossal structures are placed in an anatomically correct position, though a more detailed look reveals discontinuity in the ribs. The lung vasculature extends physiologically to the thoracic wall and distinct hilar densities are present.

There are several notable extra-thoracic elements that are and several that are not included in the generated images. Side indicators and text additions are visible in a selection of the generated images. The side indicators, usually an 'L' for the left side, are generally placed in the correct position with respect to the anatomy of the patient. Occasionally multiple similar or even different side indicators are generated for a single image. Text such as 'AP' and 'Portable' is also generated with some frequency. The text for the side indicators and the text additions is not always completely legible.

Smaller structures such as ECG leads, tubes, implants and potentially small pathology are generally absent from the generated images. Instead small attempted generations of such objects are occasionally seen. These cannot, however, be distinctly identified as such. Segal et al.[108] proposed that the progressive growing nature of the PGGAN architecture is potentially responsible for the absence of small scale structures in the generated images. Given their small size these structures would only be generated towards the end of the training process.

In the generation of images the effects of a higher truncation threshold seem to be to push the model towards making more risky generations. These generally include even more discontinuity in the ribs, abstract placement of the CXR with respect to the image frame and extensive obfuscation of the lungs. Similar effects were noted in different image domains by Shmelkov et al. [128] who noted a distinct trade-off between image fidelity and image variety. A similar trade-off is noticed in our generated images.

1) Radiologist Turing Test: We recruited two residents with one and five years of experience respectively to rate a selection of generated and real CXR images as 'real' or 'generated'. The images were presented in a 3 x 2 grid in a slideshow where at least two and at most four images were generated. The remainder per slide were real images. All images used in the evaluation were sampled at random from either the pool of generated of the pool of real images. The resolution of the screen used (2056x1329) meant that images were shown at a practical resolution of approximately 600x600 pixels. Participants were given unlimited time to evaluate each image.

Each participant rated thirty images (15 real, 15 generated). The real images were identified as such 77% of the time

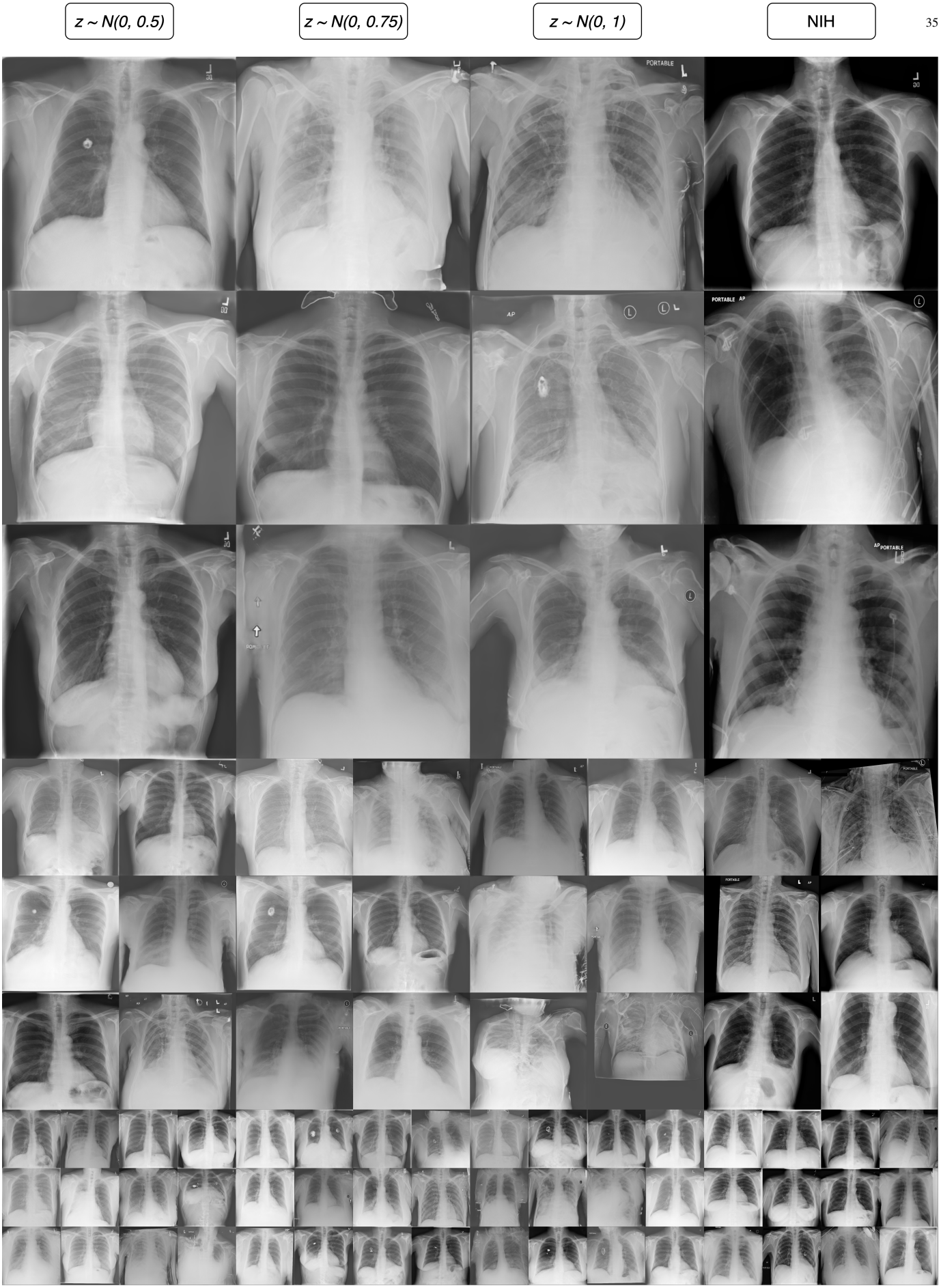


Fig. 23: Randomly generated examples by the trained generator, shown at three different resolutions. Both the generated images and the images from the NIH dataset have a native resolution of 1024x1024. Each column represents a different truncation of the normal distribution used for the sampling of the latent variable space. The right-most column is a random sample of images from the NIH dataset used to train the PGGAN.

whereas the generated images were identified as real 63% of the time. The participants mentioned that identifying generated images was primarily possible through the incoherent detail of bone structures and the incomplete generation of 'L' patient orientation markers.

2) Creating 'optimal' CXRs: Using the '14-way' image classification model from the previous chapter we've 'reversed' the generator in the PGGAN architecture by allowing an image generation towards specific disease classes. For each of the disease classes in the NIH dataset three generated examples are shown in figure 24. As was the case in figure 23 there is a normal variety in the pathology present. For a number of disease classes very evident signs of that disease class are present. Such is the case for the 'cardiomegaly' class, where an enlarged heart is clearly visible. Additionally, the 'mass' class clearly shows masses in the lungs.

There are, however, also a number of classes in which disease presence is either incomplete or very generalised. The 'pneumothorax' class shows a resemblance of a sharp delineation between what is supposed to be free air and lung vasculature, but the lung vasculature continues in the free air space. The classes 'infiltration', 'edema', 'pneumonia', 'consolidation' and 'effusion' all present very similarly with veiling of the lung fields. This is, however, in accordance with their regular presentation. Other authors have critiqued the separate inclusion of these classes in the NIH dataset for this exact reason [71, 102].

One of the more difficult disease classes is the 'nodule' class. Very few, if any, nodules are generated in these images. This links back to the earlier discussed absence of small features such as ECG leads, tubes and clips. Because of the progressive architecture it is possible that size of nodules means they are introduced 'too late' to be properly depicted.

3) Obtaining latent space representations of DRRs: In the reversal of the generator for the generation of specific disease classes we showed that it is possible to direct the generator towards a disease class specified by a discriminator. We took this a step further and tried to use a discriminator to direct the generator towards the latent space representation of an input image. Specifically, a DRR image.

In figure 25 we show the results of crudely searching the latent variable space for a representation of an input DRR image. In this representation the pose and overall look and feel of the DRR are recreated in the matching CXR. In this case there is no pathology present and therefore the recreation is passable. In more detail there are lacking characteristics such as the slightly angulated aorta knob in the DRR, which is absent in the generated image. Additionally, a letter 'L' is generated on the top right of the optimised image. This is obviously absent in the original DRR.

In an attempt to recreate a specific pathology we also looked at a case where there is a mass present in the left lung. This is shown in the second row of figure 25. This mass is not reproduced in the generated image. In fact, the generated image highly represents the generated image in the first row. It is very likely that the crude approach has led the model to attempt to create any 'good enough' visualisation of the DRR and now continually finds this again.

B. Evaluating image quality

TABLE IX: GAN-train and GAN-test performance metrics. The GAN-train metric is reported on the NIH dataset test set. The GAN-test metric is reported on the generated image test set.

Disease class	**GAN-train**	**GAN-test**
Atelectasis	$0.65(\pm 0.05)$	0.95
Cardiomegaly	$0.73(\pm 0.08)$	0.93
Effusion	$0.73(\pm 0.07)$	0.95
Infiltration	$0.61(\pm 0.05)$	0.98
Mass	$0.74(\pm 0.09)$	0.94
Nodule	$0.62(\pm 0.08)$	0.94
Pneumonia	$0.58(\pm 0.06)$	0.97
Pneumothorax	$0.61(\pm 0.05)$	0.93
Consolidation	$0.70(\pm 0.06)$	0.99
Edema	$0.71(\pm 0.07)$	0.98
Emphysema	$0.75(\pm 0.07)$	0.93
Fibrosis	$0.63(\pm 0.05)$	0.92
Pleural thickening	$0.58(\pm 0.07)$	0.93
Hernia	$0.56(\pm 0.09)$	0.92

To assess the quality of the generated CXRs in a quantitative manner we applied the GAN-train approach and trained an image classification model according to the same principles as the image classification model described in the previous chapter. Using the trained PGGAN architecture we generated a dataset of the same size as the NIH dataset. Latent variable samples were generated using a truncated normal distribution (threshold 0.75), as this was shown to improve variability at the cost of some image fidelity[129].

The resulting distribution of generated labels is included in table X. The labels for the generated images were provided by an ensemble of the trained '14-way' models. The average number of labels per image in the generated set is significantly higher (5 ± 2.8) than for the original NIH dataset (1.6 ± 0.8). This is explained in part by the extreme prevalence of the 'atelectasis', 'infiltration', 'effusion' and 'consolidation' labels. As we argued earlier, these labels, save for the first, present very similarly on a CXR and thus an image classification model might simply assign all these labels whenever such an image is encountered. Furthermore, far fewer 'no finding' images are generated, which also leads to an inflation of label counts.

The prevalence of any disease class is significantly higher in the generated images dataset. This could in part be due to the labelling process where a trained image classifier was used instead of a (group of) radiologist(s). We've shown that the image classifier achieves an average AUC of up to 0.78,

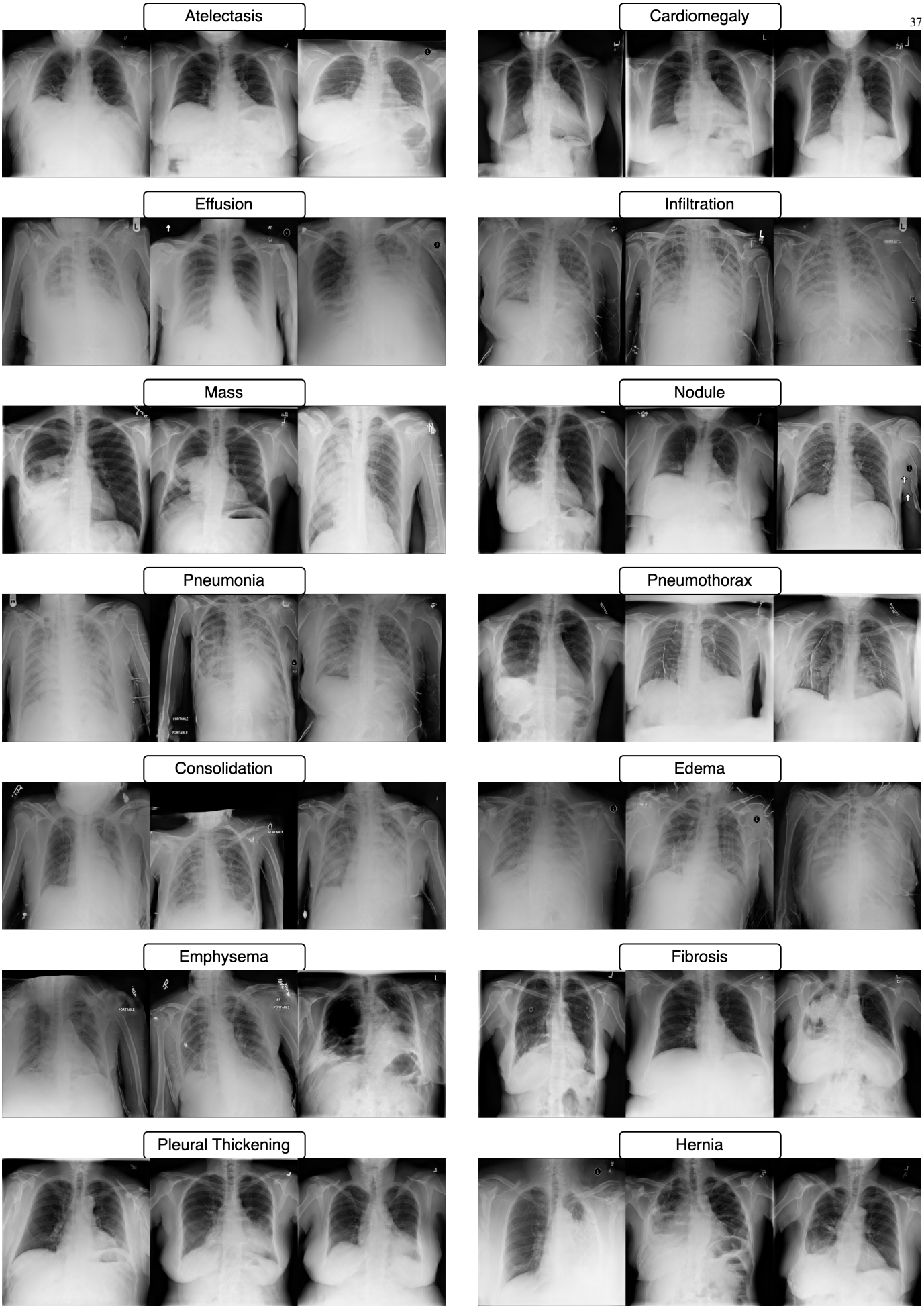


Fig. 24: Images generated using the PGGAN architecture and the '14-way' classifier model as latent space variable generation optimiser. Included are three examples of generated images for each of the classes in the NIH dataset.

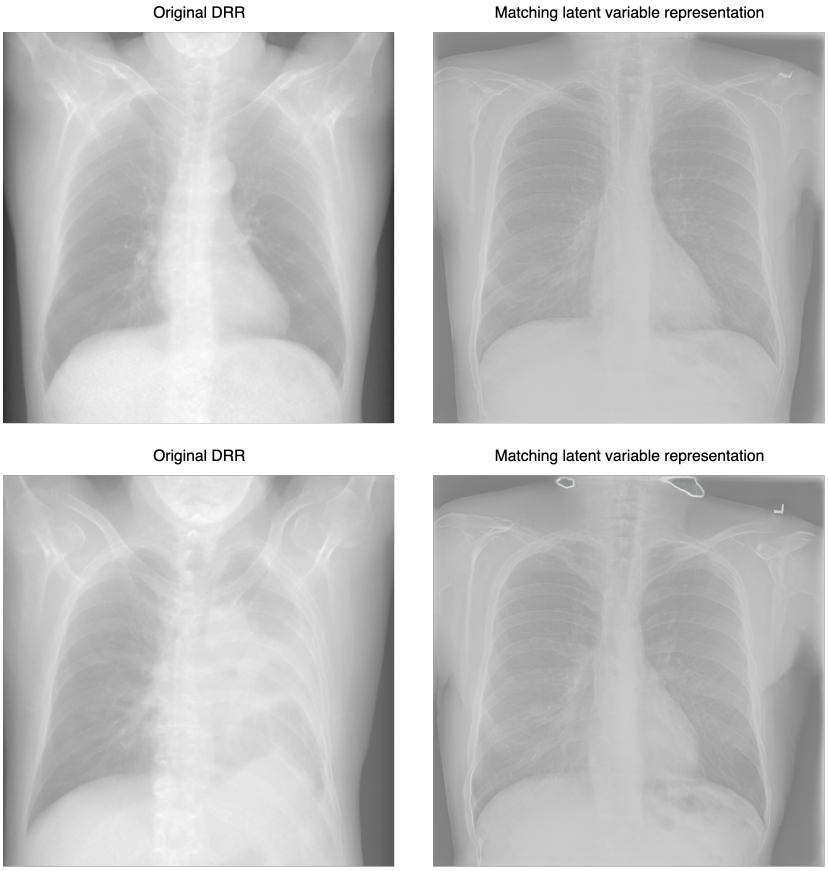


Fig. 25: Two instances of a DRR and a PGGAN generator result of the matching latent variable representation of this drr. In the top row no pathology is present, but the DRR in the bottom row contains a mass in the left lung. Both matching latent variable representations primarily represent each other and not so much the DRRs.

which implies that many regular CXRs are misclassified, let alone the generated images.

TABLE X: Comparion between the quantity of images available per disease class label between the NIH dataset and a generated dataset of roughly the same size using a trained PGGAN model.

Disease class	NIH dataset	Generated dataset
Atelectasis	11.559 (10.3%)	36.238 (34.6%)
Cardiomegaly	2.776 (2.5%)	18.196 (17.4%)
Effusion	4.667 (4.2%)	34.730 (33.1%)
Infiltration	19.894 (17.7%)	43.412 (41.42%)
Mass	5.782 (5.2%)	20.669 (19.8%)
Nodule	6.331 (5.6%)	19.262 (18.4%)
Pneumonia	1.431 (1.3%)	29.209 (27.9%)
Pneumothorax	5.302 (4.7%)	22.950 (21.9%)
Consolidation	4.667 (4.2%)	49.112 (46.9%)
Edema	2.303 (2.1%)	28.298 (27.0%)
Emphysema	2.516 (2.2%)	9.678 (9.2%)
Fibrosis	1.686 (1.5%)	15.625 (14.9%)
Pleural thickening	3.385 (3.0%)	15.085 (14.4%)
Hernia	227 (0.2%)	9.456 (9.0%)
No finding	63.016 (56.2%)	40.468 (38.6%)

This image classification model was evaluated on the NIH test set and the results for this are reported in table IX under GAN-train. The application of the '14-way' model to a test set of generated images resulted in the scores reported under GAN-test. The GAN-train evaluation shows a significantly worse performance than the '14-way' model in combination with a significantly bigger standard deviation. We believe this is in part explainable by a potential 'double-dip' of labelling errors and erroneous image generation. The '14-way' model was used to label the images on which the GAN-train model was trained, which potentially introduced errors. As we showed in figure 24, not every disease class is recreated at sufficient detail which might have contributed to the poor model performance.

The '14-way' model performance on a test set of generated images (GAN-test) showed exceptional performance. This performance does, however, come with a caveat. All predicted labels are compared to ground truth labels which were generated by an ensemble of the '14-way' models. Without a secondary independent model to label images in the GAN-test scenario these values are of limited use.

V. Discussion

The goal of this chapter was to investigate the possibilities of generating realistic CXRs and then using the underlying mechanism to apply a transformation to DRRs. In this chapter we've shown that it is possible to generate realistic CXRs using a PGGAN architecture as human evaluators rated the

generated CXR images as real more often than chance. We've also shown that it is possible to optimise images towards some discriminator output, such as a disease class. We investigated the possibilities of transforming a DRR by trying to find the latent space representation of a DRR as input image.

1) Evaluating the image quality of generated images: The CXR images generated using the PGGAN architecture are generally realistic, with a normal variation in anatomy and the pathologies present. Both male and female CXRs are created and the overall look and feel of the generated images match the original CXRs. When presented to clinical experts the real CXRs were identified as real more often than the generated CXRs, although both groups were identified as being real more often than chance. This signifies that the generated CXRs have captured the semantics of CXRs to an acceptable extent.

The application of the truncation trick [128] has the effect of creating more 'risky' visualisations as the threshold is increased. This was also found in related work on CXR generation by Segal et al. [108]. Similar results were found by Moradi et al. [126] although their generated images were created at significantly lower resolutions.

In general the smaller structures suffer the most in terms of representation and generated quality. This is also seen in the application of this architecture on human faces [123]. As discussed before, we think this is due to the relatively late introduction of 'high' resolution details into the training architecture. It is possible that a more prolonged model training time with a greater emphasis on the higher resolution details could've resolved this issue.

2) Optimisation towards a disease class: In the evaluation of the generated images several disease classes were found to be extremely similar in their presentation. These all describe some form of lung opacity where clinically it is regularly impossible to distinguish between them without having knowledge of the clinical presentation of the patient. This phenomenon also arose as a major critique point of the NIH dataset. Oakden-Rayner et al. [102] explored the NIH dataset and also found these classes to be severely overlapping to the point where a distinction could hardly be made.

It is possible that the generator has learned to create a combined average lung opacity disease image to reflect their common appearance. Potential solutions to this problem could lie in the manner with which the latent variable space is sampled. In related work by Karras et al. [110] the latent variable space is extended, which potentially allows for more control over this sampling process.

3) Creating a CXR from a DRR: In our efforts to transform a DRR into a CXR we have managed to obtain a generic CXR generator as shown in figure 25. Experimental validation shows that certain aspects such as pose and size of the thorax are largely maintained across the transformation. We think that there are evident limitations in the feature extraction approach we used to compute how closely a generated image represents an input DRR. The current approach using a VGGNET trained on an entirely different task does not suffice, as it primarily resolves DRR images down to some basic CXR form. There is no link between the presence of a pathology in the DRR and the resulting generated CXR. This approach was proposed originally by Karras et al. [111], where the application domain, i.e. real-world images, matched the application domain on which the VGGNET was trained.

Additionally, the VGGNET was trained on 224x224 images, which means that any input has to be down-sampled to be usable as input to the feature extractor. As images are generated at 1024x1024 there is inevitably a loss of semantic information. Even with a more advanced feature extractor it is possible that small details will not be captured properly as some level of randomness in the generation of a higher resolution image will persist. The StyleGAN [111] architecture might be better suited for this task. Alternatively, as was suggested by Segal et al. [108] further research would have to be done in the optimisation of latent space sampling.

4) Future work: In this chapter we've shown that it is possible to generate realistic CXR images at a high resolution. Yet this does not directly improve the image quality of DRRs generated from CT data. The StyleGAN architecture by Karras et al.[111] aims to do just that; take an input image and 'map' the image style of a second image onto that input image. It is very interesting to investigate the possibility of mapping the CXR style onto a target DRR image. Additionally, it may be possible to map variables such as patient gender, age and pathologies onto target images.

VI. Conclusion

In this chapter we've shown that it is possible to generate realistic CXR images. We've shown various examples of a natural variation in pathology, sex and anatomy. We also investigated the potential of using the CXR generation architecture to find a matching CXR for a given DRR input image. While some challenges remain, we believe that the promising results warrant a further investigation into the applicability of this approach in DRR visualisation.

Enhancing Digitally Reconstructed Radiographs with Super Resolution

I. Introduction

Super Resolution (SR) is a technical application in which a High-Resolution (HR) image is recovered from one or more Low-Resolution (LR) images. In recent years this field has attracted a lot of attention both in general imaging [130] and medical imaging research [131], with a focus on the so-called Single Image Super Resolution (SISR) where only a single LR image is used to reconstruct a HR target. SISR can be realised with something as simple as bilinear interpolation, though such a basic approach is guaranteed to blur the images despite improving the available resolution. Image super resolution is therefore generally considered a task of improving image fidelity in a given image, in which the resolution is the vessel to achieve this. In other words, the spatial resolution has to be improved.

SISR was traditionally achieved using classical approaches such as edge sharpening [132], deconvolution [133] and example based methods [134]. The introduction of a Convolutional Neural Network-based architecture by Dong et al. [135] marked the beginning of the rise of DL-driven models for this task. Such models have since rapidly developed focusing either on a Peak Signal-to-Noise Ratio (PSNR)- [136, 137] or perceptual-driven [130, 138, 139] approach. Though easy to compute, the PSNR metric has been shown to disagree with the subjective evaluation of human observers [140]. In the perceptual-driven approaches loss functions have been more difficult to define, but these approaches have led to State-of-the-Art (SOTA) results [130].

In the previous chapter we saw how AI-driven models trained on Chest Radiographs could be successfully applied to Digitally Reconstructed Radiographs. We also showed the potential of using Generative Adversarial Networks in the generation and optimisation of certain visualisations. In this chapter we aim to leverage the potential of Deep Learning based solutions to tackle some of the feedback from our initial reader study. One key piece of feedback from this reader study presented in Chapter 1 was related to the limited level of resolution in the constructed DRR images. This feedback point is directly related to the ULDCT modality from which the DRRs were constructed.

The resolution of the DRR is limited by the matrix size of the CT scanner, which is fixed at a size of 512 by 512, and the slice thickness at which the scan was reconstructed, which is set at 1mm in the LUMC dataset. The axial in-plane pixel size can be used in combination with the slice thickness to compute a set isotropic voxel size. This is necessary to avoid a non-affine transformation of the scanned images. To increase the available resolution post-processing of images is necessary; the scanner cannot physically output a higher-resolution image.

In this chapter we investigate the applicability of current existing State-of-the-Art methods on (medical) image super resolution to CXRs and DRRs constructed from ULDCT data. To the best of our knowledge no such evaluation using ULDCT data exists. We also implement and apply the ESRGAN architecture to the super resolution task to train a domain specific super resolution model. This leads to the following research question:

> *To what extent can super resolution models boost the perceived quality of DRRs constructed from ULDCT data?*

This chapter is structured in the following way. In section II we discuss the background and related work on image super resolution. In section III we describe how the image super resolution model is trained and propose a number of evaluation methods. We report the results for these evaluations and place them in context with literature in IV. We present a discussion on future work in section V and provide a conclusion in section VI. For the sake of brevity key principles related to AI models and GANs are not repeated in this chapter. This too applies to the discussion on the used datasets.

II. Background & Related Work

In this background we provide a brief introduction on the topic of super resolution. We then continue to describe the application of super resolution in (medical) related work.

A. Image super resolution

Super Resolution is the method with which the spatial resolution of a target image is improved. This is distinctly different from, for example, the improvement of image capturing equipment to obtain a higher resolution image. In that case the spatial resolution is improved primarily through improvements to the detector used in capturing the image. There are many applications in which this is unfeasible because of design limitations, undesirable due to cost or even impossible due to chip size limitations [141].

The application of SR is a favourable alternative to improving the image capturing equipment to boost image quality. The task of Super Resolution can be split into two sub-tasks:

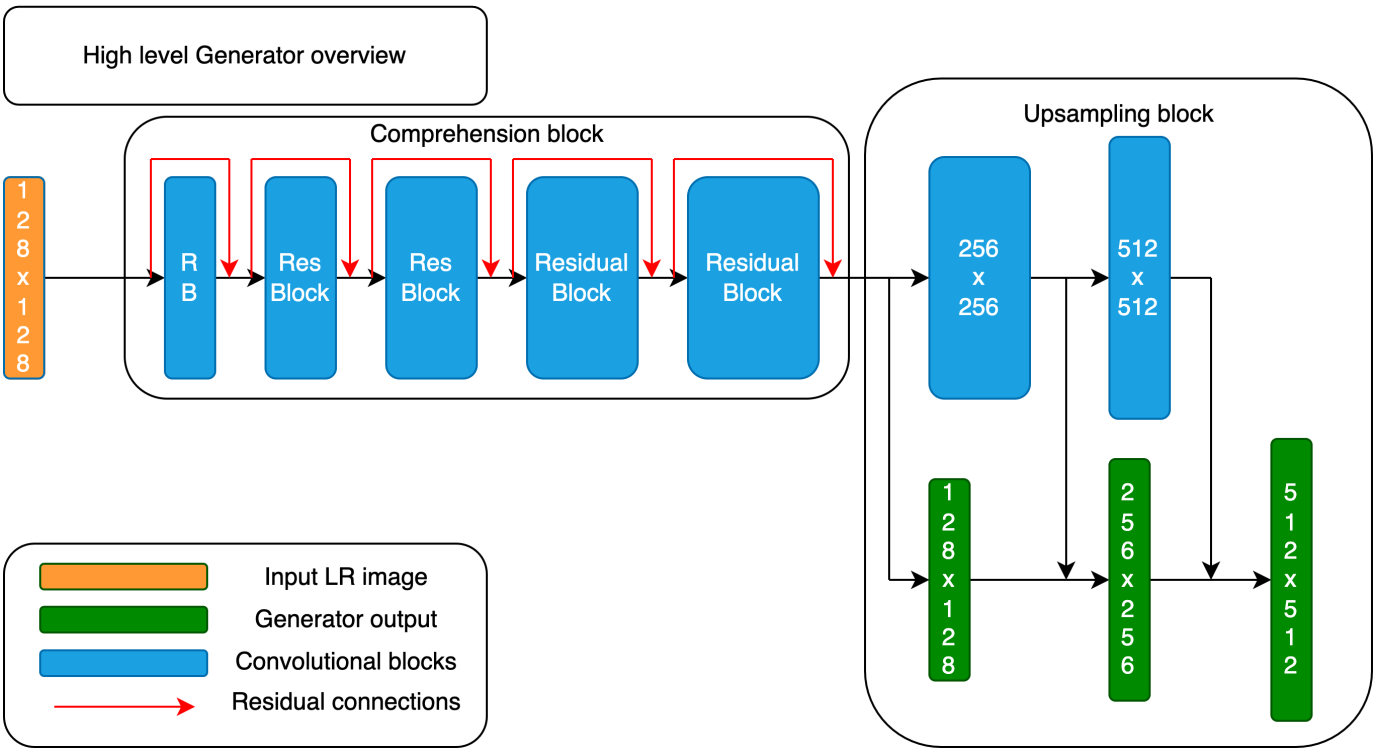


Fig. 26: High level overview of the generator of the ESRGAN architecture. The generator takes a LR image as input and forwards this through a number of residual blocks in a 'comprehension block'. Afterwards the 4x up-sampling is realised through two succesive layers of up-sampling convolutional blocks. The width of these blocks is a visual indicator of the number of convolutional layers used, i.e. the higher resolution blocks use fewer layers than their lower resolution counterparts.

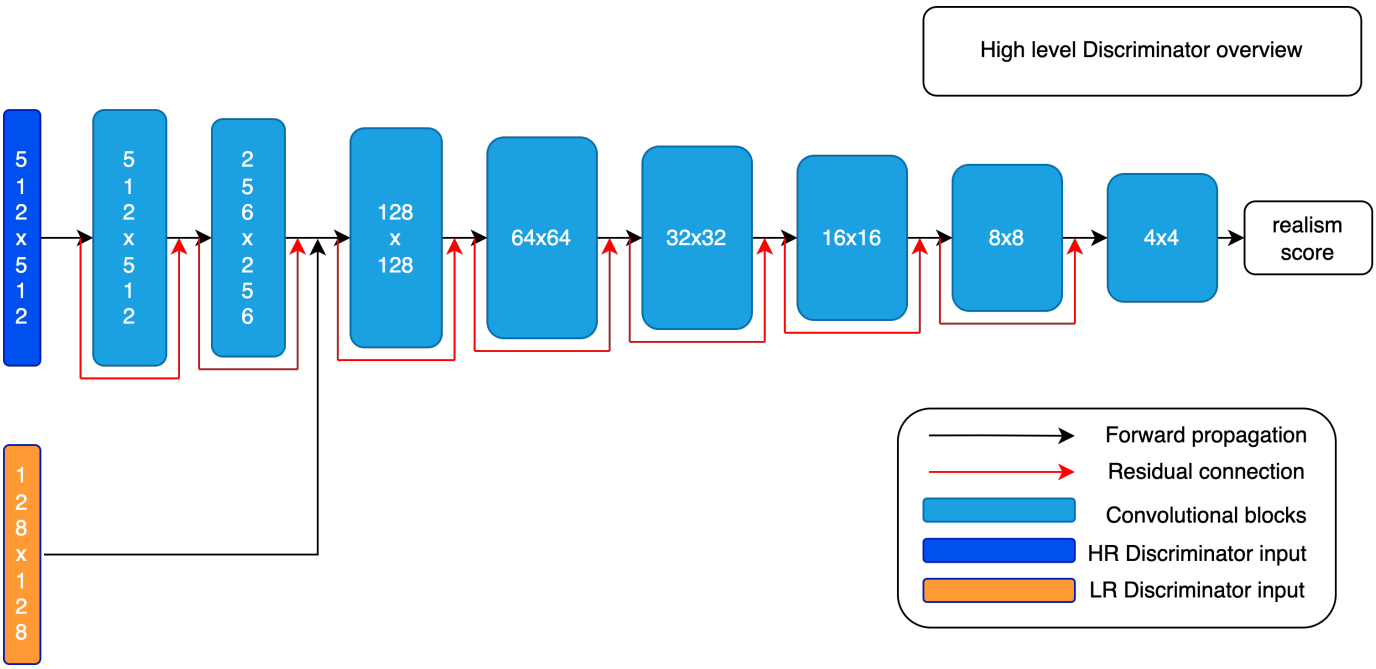


Fig. 27: High level overview of the discriminator of the ESRGAN architecture which computes a realism score for a given input image. The discriminator takes a LR image and its corresponding HR image as input. Successive layers of convolutional blocks reshape and downsample the image where the lower resolution layers use more convolutional layers per block than the initial higher resolution layers. Shown in red are the characteristic residual connections of the ResNet [79] architecture which 'skip' the convolutional blocks.

Single Image Super Resolution (SISR) and Multiple Image Super Resolution (MISR). In SISR a High-Resolution (HR) image or image patch is to be recreated from a matching Low-Resolution (LR) variant. For MISR there are a number of LR images available which are then to be combined into a single HR image.

The area of SISR has been the main focus of research into super resolution algorithms [142]. This is in part because there are not always multiple images of a same scene available, and most importantly because even if there are, SISR would in any case be able to solve those problems that MISR is able to solve. Datasets for SISR are also considerably easier to construct than for MISR.

Super Resolution has traditionally been achieved using either learning based approaches such as example-learning [134] or pixel-based methods [143] or reconstruction based approaches such as edge sharpening [132] and deconvolution [133]. The advent of the Convolutional Neural Network and its application in SR has led to these traditional methods being supplanted by supervised DL-based approaches. Even though unsupervised approaches exist, these have failed to meet the same image fidelity standard as the supervised approaches [144, 145].

The first application of a CNN to a SR task was the [135], who formulated SR as an end-to-end task. In their work features are extracted from up-sampled LR patches to find the best mapping to a corresponding HR patch. By using only convolutiona; layers this work was revolutionary in that it enabled the application of a SR model to a LR patch of any dimension.

Improvements to this approach were suggested by Lim et al. [146] who added residual image comprehension blocks to the fully connected architecture in their EDSR model. Zhang et al. [147] investigated the applicability of a residual dense block instead in their RDN SR model, reporting favourable results. A further iteration in the CARN architecture was realised by adding cascading connections into the residual blocks [148].

Perceptual-loss based approached have also been suggested for the task of Super Resolution to boost the visually perceived quality of results. Such approaches have been dominated by GAN-based architectures. State-of-the-Art results were shown by the SRGAN [140] architecture and improved by the ESRGAN [130] architecture. In both approaches photo-realism is obtained by carefully training the SR model on large datasets in combination with a perceptual or contextual loss function.

1) The Enhanced Super Resolution GAN architecture: In this chapter we're applying the Enhanced Super Resolution GAN network by Wang et al. [130] to the task of achieving SR in CXR and DRR images. This architecture has shown to be able to achieve State-of-the-Art results in SR. Because this model is a GAN, the network architecture is very similar to the network architecture of the PGGAN model used in chapter 3. The major differences between the generator used in the PGGAN architecture and the one used here are that a LR image is used as input instead of noise, and the addition of a 'comprehension block'. The generator is shown schematically in figure 26. For the discriminator both the LR as well as the HR produced by either the generator or the ground truth dataset are used as inputs to compute a realism score.

2) Evaluating SR models: The accuracy of reconstructed SR images is usually evaluated using the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) evaluation metrics [130, 149]. The PSNR is a metric that describes the ratio between the power of a reference HR image R and the power of a downgraded target LR image T. The PSNR is computed using the MSE, which is defined by:

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (R(i,j) - T(i,j))^2 \tag{14}$$

with R and T the reference and target image respectively. The PSNR is then obtained using the following equation:

$$PSNR = 10 \log_{10} \frac{(L-1)^2}{MSE} \tag{15}$$

where L represents the maximum intensity level. In a PNG image this is 255. The PSNR can be used to quantitatively evaluate to what extent a SR patch matches the original HR patch.

The Structural Similarity Index Measure is a metric that was proposed to evaluate the perceived similarity between a reference and a target image [150, 151]. In this metric the luminance, contrast and structure of the reference and target are compared and combined into one metric:

$$SSIM(R,T) = \frac{(2\mu_R\mu_T + c_1)(2\sigma_{RT} + c_2)}{(\mu_R^2 + \mu_T^2 + c_1)(\sigma_R^2 + \sigma_T^2 + c_2)} \tag{16}$$

with μ_R, μ_T the average of R and T respectively, σ_R^2, σ_T^2 the variance of R and T respectively and σ_{RT} the covariance of R and T. The variables c_1 and c_2 are used as division stabilisers based on the maximum intensity level L. Using the PSNR and the SSIM a quantitative measure of the reconstruction accuracy of a SR model can be determined.

3) Super resolution in medical related work: Super Resolution models have also been applied to medical images [142] such as Chest Radiograph SR [152], CT SR [153] and MRI SR [154]. In these applications there are significant stakes related to model performance. If the SR model does not enhance a lesion properly it could lead to a missed or altered diagnosis. At the same time, the application of SR models has the potential to alleviate physical CT or MR scanner limitations in terms of their matrix size.

One of the big remaining challenges is to find an evaluation metric that represents the perceptual perceived quality well.

As described above, performance is often reported using the PSNR and SSIM metrics, but Ledig et al. [140] showed that the PSNR metric does not always agree with the subjective evaluation of human observers. This is especially relevant in a medical context where the addition, or deletion, of crucial details can have significant consequences. The value of a human reading of SR results can for now not yet be beaten.

III. Methods

In this section we present our approach on applying SR to DRRs. As we saw in section II, a supervised model for SR primarily needs two things; vast quantities of data and corresponding LR and HR images for every entry. For DRRs we have neither. At best we can down-sample the existing DRR images and use the original DRRs as ground truth, but we can then only reasonably expect to match the resolution and not surpass it. We therefore need to train or obtain a SR model that has been trained on a very similar imaging domain (the CXR) in order to apply SR to the DRRs. In this section we discuss how we train, obtain and compare different SR models.

A. Super resolution of chest X-Rays

We implemented the Enhanced Super Resolution GAN (ESRGAN) model in TensorFlow based on the open-source code provided by Wang et al. [130][1]. This architecture incorporates a novel Residual-in-Residual Dense Block (RDDB) into the generator compared to the Super Resolution GAN (SRGAN) architecture [140] architecture. The authors also removed the Batch Normalisation layers.

The full NIH dataset was converted into TFRecords and hosted in a Google Cloud storage instance for data access. TFRecords were generated at 512x512 and 1024x1024 (native) resolution to be able to sample crops efficiently at varying resolutions during model training. Due to limitations in compiling certain TF functions on the TPU hardware the different resolution levels are provided manually. We trained our instance of the ESRGAN architecture using a combination of 128x128 LR and 512x512 HR patches. The primary reason we did not use a larger LR and HR input patch was a memory limitation in the most commonly available TPU hardware through Google Colab.

The HR patches are obtained using a random 512x512 crop on the 1024x1024 input data. This random crop is subsequently downsampled using bicubic interpolation [127] to obtain the 128x128 LR patch. This is shown schematically in figure 28.

The model is thus trained to achieve a 4x super resolution up-sampling. We made use of the discriminator to compute a content loss based on the softpluss loss function. This has been shown to provide better convergence at higher resolutions [124]. Additionally, pixel normalisation, minibatch discrimination and lazy regularisation (once every 8 steps) were applied.

We trained the model using a distributed training strategy on hardware available through a Google Colab instance consisting of 8 TPUs. We trained the model by showing 512 images per epoch to each distributed training instance until 26 million images were shown. Model training was visually inspected using a custom visualisation callback that smoothed over visualisation weights by applying weight decay and sampled images at resolutions between 512x512 and 1024x1024 to validate performance. Total model training took 3 days.

B. Clinical reader study 2: Assessing SR image quality of DRRs

The PSNR and SSIM evaluation metrics for SR images both represent a measure of the reconstruction accuracy. These do not capture the (human) perceived image quality. The PSNR was actually shown to be inversely correlated to human observer quality assessment [140, 149]. This leaves human observers as the best method to obtain an objective assessment of SR image quality.

To realise this, we conducted a follow-up clinical reader study using DRRs constructed using the DRR construction method 'softMip' [48] to evaluate the image quality of SR images. This DRR construction method was chosen due to its preference amongst the participants of the first reader study. The DRRs were produced at an isotropic voxel size for twenty cases known to not have major pathology. This was done to ensure comparability to the initial clinical reader study. The DRRs were up-sampled using our SR model to obtain a 4x base resolution. The images were saved to a DICOM format and then displayed in the DICOM viewer MicroDicom[2].

In this follow-up reader study we recruited participants to rate a selection of DRR images using a number of questions. In this selection half the images had SR applied to them and the other half didn't. The participants were not made aware in advance which images had SR applied to them. We asked experts written informed consent with regards to their participation and the sharing of aggregate personal information. In this follow-up evaluation we used the same questions from the initial reader study, again to ensure comparability. The original Dutch and translated English questions are included in chapter 2.

Due to the limited number of participants in this clinical reader study, we again also focused on the qualitative feedback provided by the participants. The feedback was analysed in depth, and by using inductive category development as described by Mayring et al. [59] open issues were identified

[1] https://github.com/xinntao/ESRGAN

[2] https://www.microdicom.com

and used in future improvements [60].

1) Participants: In order to assess the quality of SR images, we need expert domain knowledge. We recruited 2 medical professionals to participate in the second clinical reader study. At the time of participation all professionals were employed by the LUMC hospital. The two participants were both residents at the time of the study with 1 and 5 years of experience respectively.

IV. Results

In this section we present the results of the application of our SR model to CXR images and subsequently DRR images to show that cross domain application is feasible. We also present the results of our human observer evaluation study.

A. Image super resolution

We present a number of CXR images from the LUMC dataset to show the qualitative results of our super resolution model in comparison with the State-of-the-Art models RDN [147], RCAN [148], EDSR [146] and the original ESRGAN [130] super resolution models. These models are applied to select LR patches which were down-sampled using bicubic interpolation from HR crops from original images in the LUMC and NIH datasets. These images are not from the same dataset that the super resolution model was trained on.

We also report the average PSNR and SSIM reconstruction metrics on 100 CXRs sampled at random from the LUMC, NIH and LIDC-IDRI datasets in table XI. We report these scores for the different SR models and apply these at both 128x128 and 256x256 resolution for the LR patches. In all iterations our approach achieves the highest average PSNR metric if the bicubic metric is discarded. This metric is likely elevated in all images because on average, bicubic interpolation generates an image that highly resembles the original HR patch. Because this is rewarded significantly in the computation of the PSNR metric we include the bicubic score only for reference. The PSNR of our approach increase if the input LR patch is increased from 128x128 to 256x256 resolution. This suggest that the model has learned an appropriate scale invariance as a result of the mixed resolution inputs during training.

For the SSIM we find varying results. The RDN, RCAN and EDSR approaches score very similarly across all datasets. Because these approaches are not driven by a GAN-based architecture there is a reduced amount of 'filling in' data. We expect that this explains the higher SSIM scores for these approaches compared to both ESRGAN approaches we report on. This is evident in the original ESRGAN model that generally outperforms these other approaches in terms of PSNR, but achieves a significantly worse SSIM. Notably, increasing the resolution of input patches from 128x128 that the model was trained on to 256x256 improves the SSIM score in all circumstances. We believe that as the input image now contains more detail, there is less need for a SR to come up with details that don't exist and therefore achieve a higher SSIM.

We confirm in figure 29 that our domain-trained ESRGAN approach (SR) is able to outperform the other approaches in terms of edge sharpness and image-likeness when compared to the original HR patch. Additionally, our approach achieves the highest PSNR across all image crops save the bicubic examples.

The SSIM or our approach is the highest in the first two rows, but not in the last two rows. This appears to indicate that the model is 'filling-in' details in those rows that are not there in the original HR patch. This is most evidently visible in the third row. This row contains a number of tubes and lines which present poorly in the LR patch. The lumen of the line running from the bottom center to the top right of

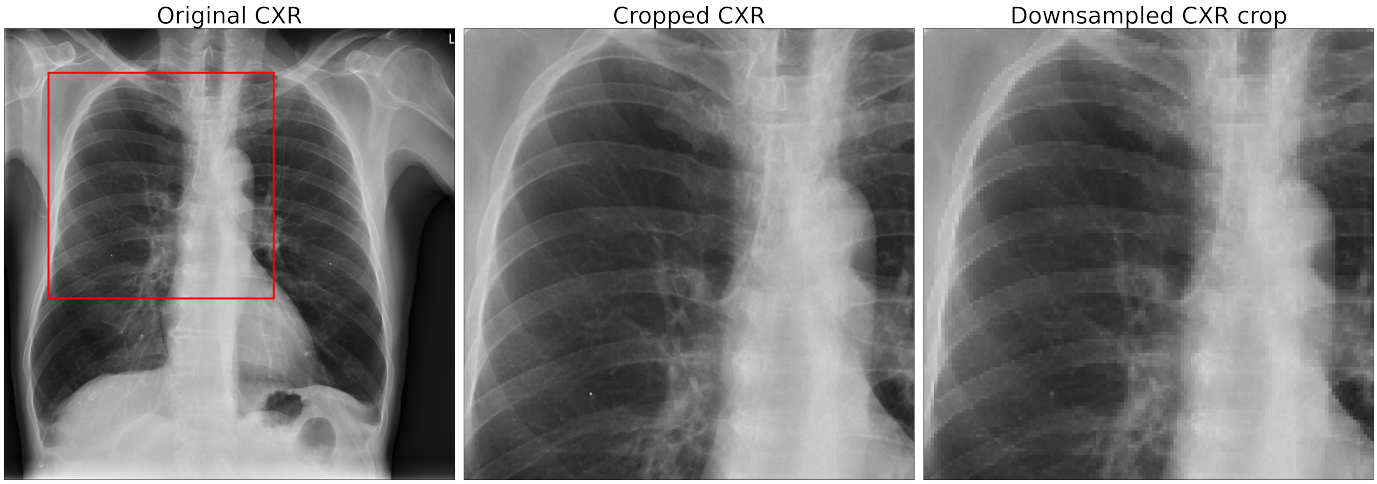


Fig. 28: Example cropping of an input image for the training of the SR model. A random 512x512 crop is obtained from an input image, denoted by the red square. This is then down-sampled using bicubic interpolation to a 128x128 image to be used as input to the generator and discriminator.

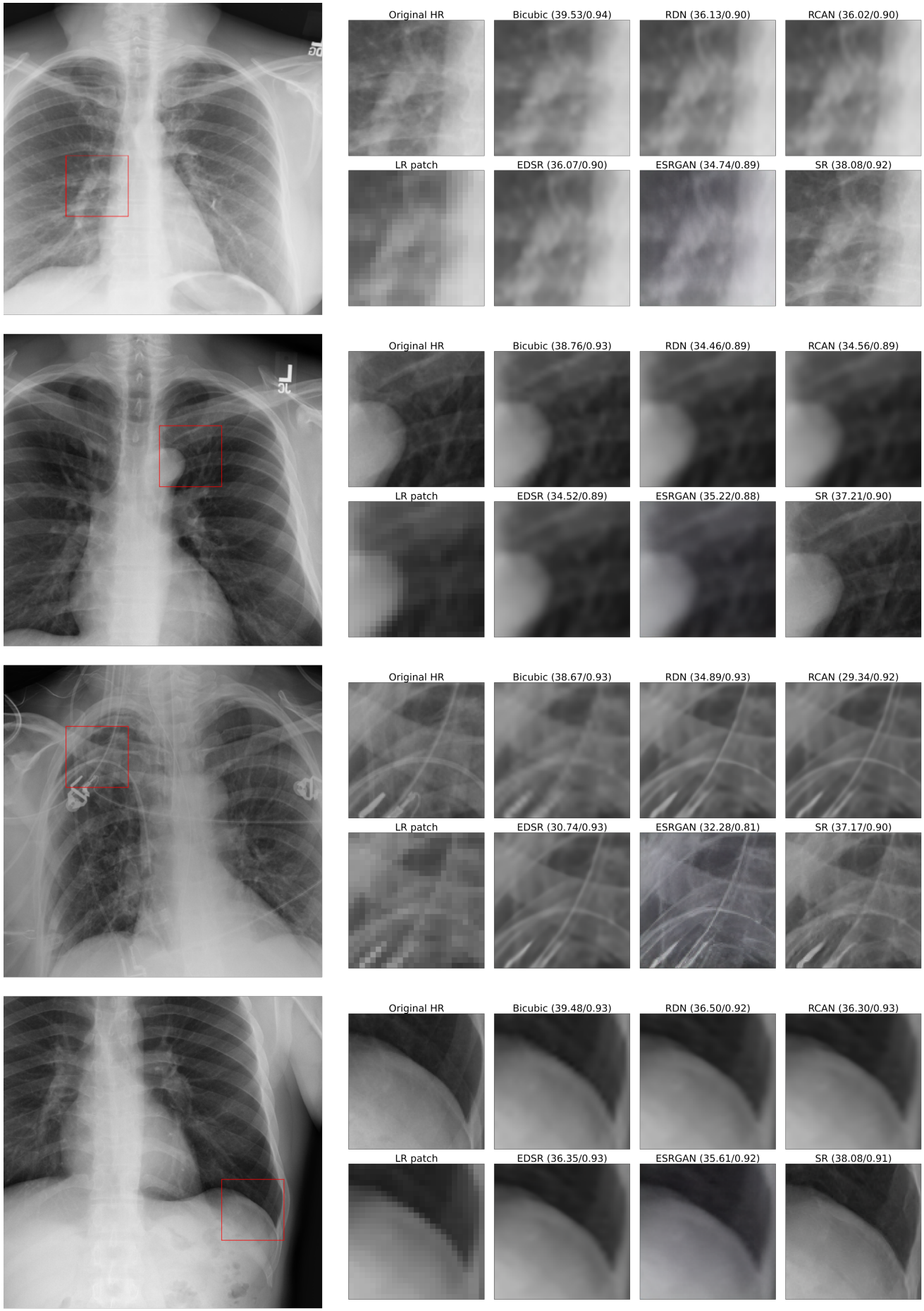


Fig. 29: Image patch super resolution results on crops of several different sizes. Reported here are PSNR and SSIM per image for a number of different super resolution architectures. Our algorithm (SR) is able to accurately recreate a variety of anatomical details.

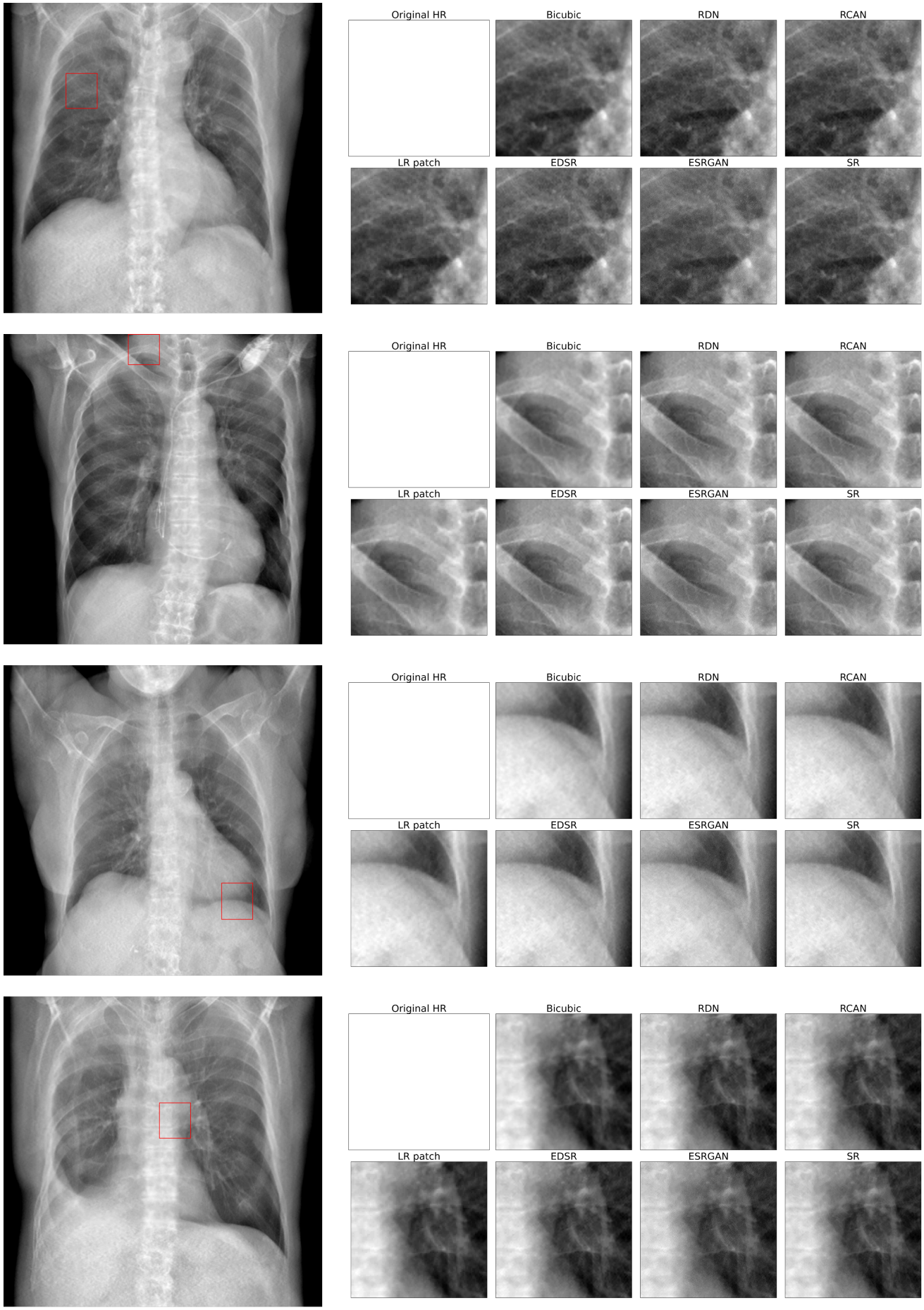


Fig. 30: Image patch super resolution results on crops of several different DRRs. The original DRRs are generated at a resolution varying between (400-580)x512 pixels.

TABLE XI: Results comparing the average PSNR and SSIM SR reconstruction metrics for a selection of 100 CXRs up-sampled with various SR models. We report results for CXRs from the LUMC, NIH and LIDC-IDRI datasets. All HR target images are re-sampled to a 512x512 resolution before being down-sampled using bicubic interpolation to a target 128x128 size. For each row the highest scores are bolded.

Dataset	LR	HR	Bicubic	RDN [147]	RCAN [148]	EDSR [146]	ESRGAN [130]	ESRGAN (ours)
LUMC	128	512	34.58/0.84	29.14/**0.78**	28.63/0.77	29.22/**0.78**	29.15/0.34	**32.05**/0.73
	256	1024	36.63/0.88	28.49/**0.81**	28.50/**0.81**	28.43/**0.81**	29.04/0.41	**33.11**/0.76
NIH	128	512	37.26/0.91	28.58/0.82	28.52/0.82	28.78/**0.83**	30.03/0.43	**34.08**/0.82
	256	1024	40.24/0.94	28.77/**0.86**	29.05/**0.86**	29.16/**0.86**	31.08/0.54	**35.85/0.86**
LIDC-IDRI	128	512	35.86/0.85	28.72/0.78	28.90/0.78	28.95/**0.79**	29.83/0.40	**32.89**/0.73
	256	1024	37.65/0.88	28.97/0.82	28.97/0.82	29.07/**0.83**	30.49/0.47	**33.83**/0.76

TABLE XII: Results from the follow-up clinical reader study. Reported are average scores from a 6-point Likert scale. In bold are the highest scores for every row.

Question	SR images	non-SR images	Initial reader study
DRR as a diagnostic CXR	**4.0**	**4.0**	3.5
Soft tissue on a DRR	**4.5**	4.0	4.4
Ossal structures on a DRR	**4.0**	3.5	3.4
Mediastinum on a DRR	**4.0**	3.5	3.9
Lungs on a DRR	**4.5**	3.5	3.4

the image is least accurately reconstructed by our approach. It is possible that these specific detailed structures are more common in the real-world dataset that the other models have been trained on compared to the relative scarcity of such images in our dataset. On the other hand, the RCAN and original ESRGAN results amplify the right border of the lumen beyond what is originally present.

In the first and second row, the model is able to recreate the vessel delineations surrounding the right hilum and the aortic knob most accurately. This is possible despite near total annihilation of the vessel detail in the LR patch. We examined the image in the first row in greater detail in figure 31. Here we show the pixel wise difference between the HR patch and the SR patch. Two areas marked with rectangles show minor differences in the recreation of vasculature and rib bone cortices. It is not clear whether these changes represent a clinically relevant alteration of the shown anatomy.

In the final row the left pleural cavity angle is the sharpest in our example, although there appears to be some deletion of the dorsal 11th rib that passes over the diaphragm in the HR patch. Where this dorsal rib meets the ventral aspect of the 9th rib there is an odd discontinuity in the cortical aspect of this rib. These irregularities most likely stem from the absence of clear information in the original LR patch. The cranial cortex of the 11th rib in this patch could also be interpreted as a vessel.

The original ESRGAN super resolution model is not

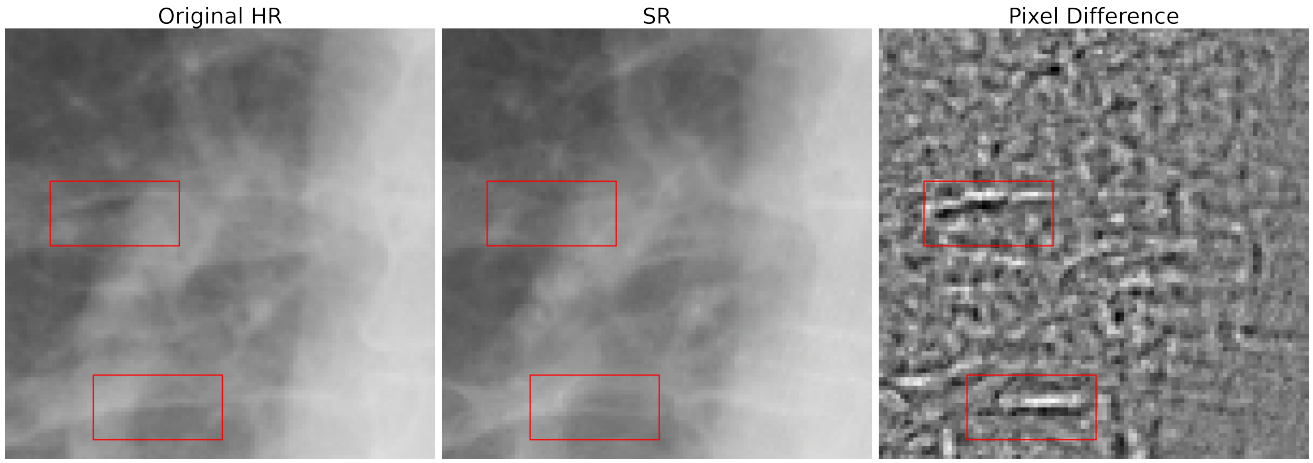


Fig. 31: The original HR image (left), the SR using our model (middle) and the pixel difference (right) for one example from figure 29. The pixel difference is expressed using positive (white) and negative (black) pixels whereas grey indicates a similar pixel value. Two areas of significant deviation are marked using red rectangles.

able to accurately capture similar edge details and image fidelity compared to our domain-specific trained version of the same model. Additionally, in all crops using the original ESRGAN there appears to be a colour shift that is not present in the other non-GAN based approaches. This could be an artefact resulting from being trained on real-world RGB images as opposed to grayscale images we used here.

B. Super resolution applied to DRRs

We also applied the ESRGAN SR model to DRR images. We show a selection of qualitative results in figure 30. Here we show the results of the 'softMip' [48] projection approach, as this was found to be the favoured approach in Chapter 2. The DRR images on the left are resized for visualisation but did not have their aspect ratio's changed. We do not report PSNR or SSIM metrics because there is no HR patch with which these can be computed.

As evidenced by the first and fourth row, our approach is able to generate the sharpest up-sampling of the LR patches without adding substantial noise. The generically trained ESRGAN model approaches similar levels of sharpness but achieves this at the cost of adding considerable quantities of noise. The RDN, RCAN and EDSR approaches don't match the same level of visual fidelity with respect to the lung vasculature and the hilar densities in the first and fourth row respectively. In the ossal structures of the second and third row our model is again able to generate the sharpest delineation, yet these improvements are at best marginal. In the abdominal structures of the third row limited improvement is visible for any of the models, but this is a very noise part of the input DRR.

We believe that the level of noise plays a considerable role in the results presented here. The DRRs are constructed from noisy ULDCT scans. This noise carries over to the reconstructed DRRs. As a result, the SR models are applied to noisy image patches. In the RDN, RCAN, EDSR and the original ESRGAN approaches this noise is still present in the up-sampled patch. In our approach, however, the model seems to try to eliminate the noise in the up-sampled patch by drawing numerous 'hair-thin' lines across all structures. This is most noticeable in the ossal structures of the second row patch and the abdominal structures in the third row patch. Such 'added' structures are undesirable even though they come at a reduction of noise in the images. We feel that an approach towards the de-noising of ULDCT scans prior to projection as a DRR is a warranted avenue to investigate further.

C. Follow-up Clinical reader study

The results from the follow-up clinical reader study are summarised in table XII. In this table we present the evaluation results for SR images, non-SR images and a comparison with the initial reader study. All DRRs were constructed using the 'softMip' approach. The participants (n=2) in the follow-up reader study were different from the participants in the initial reader study (n=6). The DRRs with SR applied to them scored higher in all categories compared to the initial reader study. The biggest difference was found for the evaluation of the lungs.

V. Discussion

The goal of this chapter was to investigate whether Super Resolution models could boost the perceived image quality of DRRs constructed from ULDCT data. In this chapter we've investigated which SOTA SR models exist and what their applicability to CXRs is. We also trained our own instance of the ESRGAN model architecture and showed that this domain-specific application achieved the highest image fidelity and restoration accuracy across a number of datasets. In this section we discuss these results and try to identify what future steps are necessary.

A. Super Resolution performance

Our SR model outperformed all other referenced models. It achieved a greater PSNR than the referenced metrics and the examined qualitative results show greater image fidelity and a more accurate and sharp recreation of a HR patch using our approach. Zhao et al. [131] also trained and applied a SR model to CXR images. In their work they reported a significantly higher PSNR and SSIM (38.14/0.93 vs our 35.85/0.86) on 4x SR applied to CXRs from (a selection of) the NIH dataset. Comparing the visual quality of their generated SR patches does not, however, explain this difference. We are of opinion that our approach achieves better quality images despite an apparent difference in PSNR and SSIM measures. We believe these differences may have been caused by the unknown image resolution at which the authors obtained their comparisons, in combination with the use of a different test set.

Xu et al. [155] also applied a GAN-based architecture to the task of SR in CXRs. They too report a higher SSIM (0.91 vs our 0.86) on the 4x super-resolution task. In their work it is also not clear at which resolutions these figures were obtained. In this comparison we take our 128x128 $->$ 512x512 up-sampling evaluation. Due to the low starting resolution there is a lot of room for our model to 'make up' details in the target image. We saw this in figure 31 where even at a LR patch of 128x128 the model has added certain structures. The addition of these will negatively impact the computed SSIM.

All of the models we referenced when comparing our SR model were trained on 'Real-World' data, that is to say that RGB images of objects such as cars, people, landscapes and buildings were used to train these SR models. Images in the medical domain are unique in that they do not have a colour channel; they are only grey. For the SR models we therefore created RGB images out of our grey-scale images to be able to use these models. This is not the expected data input, however, and we saw that the original ESRGAN model even applied a colour shift to the up-sampled patches.

The application of SR to colour images has been identified as a bigger challenge than just grey-scale images because of the complexity of the interaction between the different colour channels [156]. In this case the referenced models are potentially simply mistrained for the applied task.

As we saw in figure 31, there is a potential for details, or rather artefacts, to be generated using a LR patch that are not present in the original HR patch. This is undesirable as it opens the door the possibility of details being added that were not in the original image data. Especially in a medical context this could be very dangerous when it leads to a different or missed diagnosis.

B. Follow-up Clinical reader study

In the follow-up clinical reader study we showed that the SR model boosted the perceived quality of DRRs constructed using the 'softMip' approach when compared to images that did not have SR applied to them. This is in line with expectations although the results stem from an evaluation with several limitations. The limited number of participants (n=2) in the follow-up clinical reader study means that these results are a good indication of improved quality but don't constitute a significant result. Additionally, no cases with pathology were included to ensure comparability to the initial clinical reader study. It is not known if and to what extent pathology would be reconstructed properly with the SR model. The increase in numbers of participants and the inclusion of pathology is left as future work.

C. Future work

In the application of our SR model to the DRR images we witnessed a very interesting phenomenon where noise was removed by adding in hair-thin structures. Because the SR model was originally trained on a clean dataset, there is no learned behaviour with regards to noise in the input images. For this a separate de-noising model, or the addition or more noisy data in the training set could see a performance improvement.

In Chapter 2, we found in our first reader study that the levels of noise in the constructed DRRs were noticeable and degraded the perceived image quality. This further supports our belief that de-noising is a warranted approach to investigate. In recent work, there has been an investigation into the potential of combining de-noising and Super Resolution [157], although this approach added Gaussian rather than CT-specific noise to the target datasets.

VI. Conclusion

In this chapter we presented the application of a State-of-the-Art Super Resolution architecture to Chest Radiographs and Digitally Reconstructed Radiographs. We showed through a qualitative evaluation that a model trained on CXRs can boost the perceived image quality in DRRs. We further confirmed this in our follow-up clinical reader study, where the SR DRRs consistently scored better than the non-SR counterparts. We identified the noise level to be a key area of future work.

Discussion

In this general discussion chapter I summarise the results of my thesis by looking back on the research questions. I then proceed by discussing some overarching themes and set out some plans for the future.

A. Creating synthetic chest X-Rays

To investigate the synthetic chest X-Rays I posed the following research question:

> *What methods exist to generate synthetic chest X-Rays from (ULD)CT data and how are these perceived quantitatively and by clinical experts?*

In chapter 2 I found that there are four main different methods described in literature of constructing a synthetic chest X-Ray, also referred to as a Digitally Reconstructed Radiograph. I evaluated the construction methods using an automated histogram-based approach and consulted clinical experts on their opinions of the generated images. Such a comparison between different DRR construction methods had not yet been performed.

The results from both the quantitative as well as the expert evaluation suggest that there is a general preference for the 'softMip' approach by Meyer et al. [48]. In this approach the authors tried to combine the edge sharpness of a MIP and the noise suppression of an AVG image setting. Both settings are routinely used in clinical practice and their combination seems to suit DRRs constructed from ULDCT data especially well.

The DRRs were evaluated by the clinical experts in known absence of pathology. This was commonly raised as a limitation of the evaluation by participants. The level of noise especially in the abdomen and the spatial resolution of the whole image were raised as points of technical concern. Despite these shortcomings, participants generally accepted the DRR representation for a case without pathology.

B. Using AI disease classification models on chest X-Rays and Digitally Reconstructed Radiographs

Inherent limitations in being able to evaluate various projection optimisations with clinical experts led to the desire for a semantically viable automated evaluation method. I researched the applicability of AI-models to CXRs and DRRs using the following research question:

> *Can AI-models trained for chest X-Ray disease classification be used to evaluate the (diagnostic) image quality of Digitally Reconstructed Radiographs?*

In chapter 3 I showed the possibilities of applying DL models to CXRs and DRRs to both classify these according to a number of diseases, but also to see if we can quantitatively measure the effects of image processing on the input DRRs. I applied the SOTA architecture CheXNet by Rajpurkar et al. [71] to the task of image classification, where I noted a decrease in performance compared to their published results.

Nonetheless, I showed that it is possible to fine-tune and apply this model to greater success on a combined dataset of DRR and CXR images. This finding is in line with work by Mortani Barbosa et al. [32] who also trained a joint CXR and DRR model specifically for the detection of COVID-19. Out of fourteen disease classes that are distinguished between those that are small in size or detailed by nature such as nodules suffered the worst classification performance. This was also found in the work of Segal et al. [108] who noted that the resolution at which the images are fed into the model (224x224 pixels) potentially obfuscates the detail required to classify nodules. A validation of this on an external dataset of only nodules proved difficult, as I found these images to be highly susceptible to small alterations with regards to classification performance.

Having established that AI-models can successfully be applied to DRRs, I then investigated to what extent such models can be used as a quality evaluator of various image post processing techniques. Amongst these were variations in window width and window level settings and histogram equalisation based methods. I showed that adjusting window width and window level settings in addition to applying histogram equalisation generally improved disease classification performance in DRRs, but degraded the performance in regular CXRs. Li et al. [158] and Salem et al. [61] too showed an improvement in AI-model classification performance in altering window width and window level settings and applying histogram equalisation respectively. This further solidified this approach as being viable in determining image quality automatically.

C. Generating realistic chest X-Rays and the implications for Digitally Reconstructed Radiographs

Despite being able to automatically evaluate the effects of different image post processing techniques on DRRs, this did not help me find the 'optimal' visualisation for a given DRR. In order to actually represent a DRR optimally as a CXR I had to create a model that could generate realistic CXRs. This led me to the following research question:

Can AI models be used to generate realistic CXR and can they subsequently facilitate the generation of a CXR visualisation for a DRR?

In chapter 4 I investigated whether it was possible to generate an 'optimal' image with respect to a AI-model or DRR. For this I implemented a GAN-based CXR generation model. With this model I showed that it is possible to generate realistic CXR images. I recruited experts to evaluate the realism of generated CXRs, were even though real CXRs were identified as such more often than the generated CXRs, both occurred more than 50% of the time.

I also demonstrated that by replacing the discriminator in the GAN architecture with a disease classification model it is possible to generate images of specific disease classes. Similar results were achieved using varying architectures in the work by Segal et al. [108] and Moradi et al. [126]. In my work novel alterations were made regarding the network architecture and the resolution images were generated at respectively.

I continued the research into the generation of CXR images by investigating whether it was possible to 'invert' the generator of the GAN architecture and thus obtain a latent vector representation of a DRR. By using a VGGNET as feature extractor like in the work of Karras et al. [111] this enabled me to create a CXR-like representation for a target DRR. This is a novel contribution to the field of GAN-based CXR generated images. It does not, however, constitute a major breakthrough as I found that most matched generated CXR images were very much alike and did not represent pathology as it was present in a target DRR. Yet it remains an interesting direction to pursue for future research.

D. Applying Super Resolution to Digitally Reconstructed Radiographs

One of the key feedback points raised by experts recruited for the first evaluation was the resolution of constructed DRRs. This led to the following final research question:

To what extent can super resolution models boost the perceived quality of DRRs constructed from ULDCT data?

Super Resolution models generally require large quantities of high resolution target images from which low resolution source images can be constructed. For DRRs I had neither. In chapter 5 I therefore implemented a SOTA Super Resolution model architecture and trained this specifically on CXR images. This resulted in SOTA results in PSNR and SSIM metrics, comparable to the work of Wang et al. [130]. I made the novel alteration of using the discriminator as realism score for the generator.

I then showed that it is possible to apply this model to DRR images as well, by recruiting experts and having them evaluate low and high resolution DRR images. Here I reported a consistent improvement in their valuation of images that had SR applied to them compared to those that did not. Such a similar evaluation across imaging domains had not yet been performed, and showed the potential of applying SR to a closely linked yet different domain.

E. General dataset limitations

In this work I made use of several publicly available datasets as well as a dataset collected in the LUMC. None of these datasets were truly clean in their provided or created labels. The NIH dataset specifically was published with a disclaimer that up to 90% of all labels are accurate. Further investigations by Oakden-Rayner et al. [102] showed that for certain disease classes this was an overestimation. Such labelling errors have a knock-on effect on multiple aspects of the work that I've done. These are reflected in the disease classification model, although this was mitigated to some extent by the re-labelling of Rajpurkar et al. [70].

Another aspect worth consideration is the 'single'-centre collection of the data. The NIH dataset was collected in a single collective of health institutes. This is known to be vulnerable to selection biases in data representation [70] and potential issues regarding trained model generalisation on external datasets [71]. Yet at the same time this dataset is sufficiently large that such issues are mitigated. This was evidenced in part by the reasonable performance of my disease classification model on the external LIDC-IDRI dataset.

The limited size of the internal LUMC dataset both in actual number of cases as well as the pathology represented in these cases meant that comparisons with larger scale datasets fall short. The risk of inherent biases in such comparison is simply too great. Yet at the same time the dataset is varied enough to present at least some pathology to disease classification models, and make use of the inverted CXR generation pipeline. A key goal of future work should be to gather a larger dataset consisting of ULDCT scans and CXRs.

F. Image noise

In every chapter of this thesis the level of noise present in either the underlying ULDCT data or the constructed DRR played a role. In the first clinical reader study experts noted that the level of noise in the DRRs affected their perceived diagnostic quality of the images. In the third chapter we found significant differences between the differnt DRR construction methods with respects to the performance of a disease classification model. Here different levels of noise suppression based on the DRR may have affected these results. In the fourth chapter we saw how a CXR generation model could be used to find a matching CXR for a DRR. We found it difficult to find such a generated CXR, as the optimisation process could not account for the noise in the DRRs. Finally, the Super Resolution model in chapter 5 struggled especially with the more noisy aspects of the DRRs.

As a matter of future work, an investigation into the

de-noising of ULDCT scans seems warranted. Such a tool would be useful at all stages of the DRR interpretation process.

G. When is a DRR good enough?

The motivation of this thesis hinges on the proposal that a DRR can represent or act as a CXR-like representation of information contained in an ULDCT scan. In this thesis I've thoroughly focused on the methods with which a DRR can be constructed and how the quality of a DRR can be improved through post processing or super resolution techniques. The goal here has been to create as good as possible a DRR such that it can fulfil its role in the overall proposal.

With the advances I've shown in this thesis I feel the question of when it is good enough is warranted. As I showed with the generation of CXR images it is theoretically possible to create an optimal visualisation, but the question then should be when and how this new visualisation would be used in an actual clinical setting. For the intended purpose, i.e. summarising an ULDCT scan in a known CXR format, the current DRR is good enough.

Future efforts into for example de-noising are so beneficial not only to the DRR but to the quality of the ULDCT as a whole that these are interesting to pursue. But further improvements specifically to a DRR should perhaps be limited in their scope and be done as an exploratory work only. Efforts should instead be directed towards integrating DRRs into an AI-driven workflow where their intended use can be fulfilled.

Conclusion

In this thesis I've thoroughly investigated the different possible methods of creating and optimising Digitally Reconstructed Radiographs that are constructed from ULDCT data. In this process I've successfully applied AI models as disease classifiers and super resolution enablers to both CXRs and DRRs. I've shown the possibilities of using AI to create optimal DRR visualisations and I've evaluated my findings with clinical experts. As I stated in the introduction my goal was never to match or surpass the image quality of the CXR. With the improvements that I've realised I feel confident in saying that the DRR has reached a sufficient level of quality for its intended purpose.

Acknowledgements

I'd like to thank Hildo for his enthusiasm, his limitless vision and his infectious passion. I learned more about inflection points in space or Apple customer support than I ever could've imagined. I'd like to thank Merlijn for being a consistent and patient supervisor who let me have my creative and at times strange research related outbursts but managed to then subtly point me towards the right direction again. I also want to extend my thanks to all other members of the ULDCT and further teams at Philips; Jaap, Sandra, Hubrecht, Christian, Axel and others for the valuable cooperation of the past year. Here's to many more fruitful years of cooperation.

I would also like to thank Prof. Slump for his supervision and his role as the chair of my exam committee. His ability to abstract away from and deep dive into details has been a valuable asset in our conversations this past year. I also thank Mannes for his consistent presence and his focus that helped me separate what's important from a mountain of interesting topics to pursue. I extend my thanks to Ruby who has helped me see the value of being and appreciating myself. In our conversations these past two years I learned more about myself than I thought possible.

There is of course also a myriad of family and friends who I've been able to count on for support, for a listening ear to complain to and also for their distractions when I needed them. Thanks everyone.

GLOSSARY

AI Artificial Intelligence. 5, 20, 22, 23, 28, 30, 40, 50–53
ALARP As Low As Reasonably Possible. 9
AP Anteroposterior. 8
AUC Area Under Curve. 24–27, 29, 36
AVG Average Projection. 18, 19, 50

BN Batch Normalisation. 43

CAD Computed-Aided Diagnostics. 4, 5
CARN Cascading Residual Network. 42
CLAHE Contrast Limited Adaptive Histogram Equalisation. 25, 27
CNN Convolutional Neural Network. 4, 5, 21, 40, 42
CPU Central Processing Unit. 11
CT Computed Tomography. 4–6, 8–11, 14, 17–19, 21, 23, 25, 30, 32, 39, 40, 42, 49
CXR Chest Radiograph. 2–6, 8, 11, 13–30, 32–34, 36, 38–40, 42–44, 47–53

DICOM Digital Imaging and Communications in Medicine. 15, 16, 19, 23, 43
DL Deep Learning. 4, 5, 20–23, 29, 30, 32, 40, 42, 50
DRR Digitally Reconstructed Radiograph. 2–6, 10–30, 32, 33, 36, 38–40, 42–44, 46–53

ECG Electrocardiogram. 34, 36
EDSR Enhanced Deep Super Resolution. 42, 44, 47, 48
ESRGAN Enhanced Super Resolution GAN. 3, 40–44, 47, 48
eV electron Volt. 6

FBP Filtered Back-Projection. 9

GAN Generative Adversarial Network. 30, 32, 33, 36, 38, 40, 42, 44, 48, 51
GPU Graphics Processing Unit. 11, 14, 24

HR High-Resolution. 40–44, 47–49
HU Hounsfield Units. 9, 10, 12, 29

IDRI Image Database Resource Initiative. 23, 25, 27, 29, 44, 47, 51

LAC Linear Attenuation Coefficient. 7, 10
LD Low-Dose. 4
LDCT Low-Dose Computed Tomography. 4, 23, 24, 29
LIDC Lung Image Database Consortium. 23, 25, 27, 29, 44, 47, 51
LR Low-Resolution. 40–44, 47–49
LUMC Leids Universitair Medisch Centrum. 23–27, 29, 40, 44, 47, 51

MIP Maximum Intensity Projection. 18, 50
MISR Multiple Image Super Resolution. 42
ML Machine Learning. 20, 21
mm millimetre. 40
MR Magnetic Resonance. 42
MRI Magnetic Resonance Imaging. 14, 30, 42
MSE Mean Squared Error. 42
mSv milliSieverts. 9

NIH National Institutes of Health. 23, 24, 26, 27, 29, 33–39, 43, 44, 47, 48, 51
NLP Natural Language Processing. 20, 23, 24

PA Posteroanterior. 5, 6, 8, 14, 23
PACS Picture Archiving Communication System. 17, 19, 20
PGGAN Progressively Growing GAN. 31–39, 42
PNG Portable Network Graphic. 23, 25, 42
PSNR Peak Signal-to-Noise Ratio. 40, 42–45, 47, 48, 51

RCAN Residual Channel Attention Networks. 44, 47, 48
RDDB Residual-in-Residual Dense Block. 43
RDN Residual Dense Network. 42, 44, 47, 48
RGB Red Green Blue. 48
ROC Receiver Operating Characteristic. 24
RPL Radiological Path Length. 10, 11

SD Standard Deviation. 15
SISR Single Image Super Resolution. 40, 42
SOTA State-of-the-Art. 5, 40, 42, 44, 48–51
SR Super Resolution. 3, 5, 40, 42–44, 47–49, 51
SRGAN Super Resolution GAN. 42, 43
SSIM Structural Similarity Index Measure. 42–45, 47, 48, 51

TF TensorFlow. 43
TPU Tensor Processing Unit. 33, 43

ULD Ultra Low-Dose. 4–6, 9, 14
ULDCT Ultra Low-Dose Computed Tomography. 4, 5, 9, 13, 14, 17–19, 24, 40, 48, 50–54

REFERENCES

[1] L. J. Kroft, L. van der Velden, I. H. Girón, J. J. Roelofs, A. de Roos, and J. Geleijns, "Added Value of Ultra–low-dose Computed Tomography, Dose Equivalent to Chest X-Ray Radiography, for Diagnosing Chest Pathology," *Journal of Thoracic Imaging*, vol. 34, no. 3, pp. 179–186, May 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6485307/

[2] M. Tækker, B. Kristjánsdóttir, O. Graumann, C. B. Laursen, and P. I. Pietersen, "Diagnostic accuracy of low-dose and ultra-low-dose CT in detection of chest pathology: a systematic review," *Clinical Imaging*, vol. 74, pp. 139–148, Jun. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S089970712030557X

[3] S. Raoof, D. Feigin, A. Sung, S. Raoof, L. Irugulpati, and E. C. Rosenow, "Interpretation of plain chest roentgenogram," *Chest*, vol. 141, no. 2, pp. 545–558, Feb. 2012.

[4] S. Tigges, D. L. Roberts, K. H. Vydareny, and D. A. Schulman, "Routine chest radiography in a primary care setting," *Radiology*, vol. 233, no. 2, pp. 575–578, Nov. 2004.

[5] P. O. o. t. E. Union, "Medical radiation exposure of the European population." Mar.

2015, iSBN: 9789279453748 9789279453731 Publisher: Publications Office of the European Union. [Online]. Available: http://op.europa.eu/en/publication-detail/-/publication/d2c4b535-1d96-4d8c-b715-2d03fc927fc9/language-en

[6] I. A. Cowan, S. L. MacDonald, and R. A. Floyd, "Measuring and managing radiologist workload: Measuring radiologist reporting times using data from a Radiology Information System," *Journal of Medical Imaging and Radiation Oncology*, vol. 57, no. 5, pp. 558–566, 2013, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1754-9485.12092. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/1754-9485.12092

[7] E. E. Coche, B. Ghaye, J. d. Mey, and P. Duyck, Eds., *Comparative Interpretation of CT and Standard Radiography of the Chest*, ser. Diagnostic Imaging. Berlin Heidelberg: Springer-Verlag, 2011. [Online]. Available: https://www.springer.com/gp/book/9783540799412

[8] R. Smith-Bindman, J. Lipson, R. Marcus, K.-P. Kim, M. Mahesh, R. Gould, A. Berrington de González, and D. L. Miglioretti, "Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer," *Archives of Internal Medicine*, vol. 169, no. 22, pp. 2078–2086, Dec. 2009.

[9] C. Rampinelli, D. Origgi, and M. Bellomi, "Low-dose CT: technique, reading methods and image interpretation," *Cancer Imaging*, vol. 12, no. 3, pp. 548–556, Feb. 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3569671/

[10] F. J. Larke, R. L. Kruger, C. H. Cagnon, M. J. Flynn, M. M. McNitt-Gray, X. Wu, P. F. Judy, and D. D. Cody, "Estimated radiation dose associated with low-dose chest CT of average-size participants in the National Lung Screening Trial," *AJR. American journal of roentgenology*, vol. 197, no. 5, pp. 1165–1169, Nov. 2011.

[11] T. M. Svahn, T. Sjöberg, and J. C. Ast, "Dose estimation of ultra-low-dose chest CT to different sized adult patients," *European Radiology*, vol. 29, no. 8, pp. 4315–4323, Aug. 2019.

[12] S. H. Bradley, S. Abraham, M. E. Callister, A. Grice, W. T. Hamilton, R. R. Lopez, B. Shinkins, and R. D. Neal, "Sensitivity of chest X-ray for detecting lung cancer in people presenting with symptoms: a systematic review," *The British Journal of General Practice: The Journal of the Royal College of General Practitioners*, vol. 69, no. 689, pp. e827–e835, Dec. 2019.

[13] N. L. S. Trial, "Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening," *New England Journal of Medicine*, vol. 365, no. 5, pp. 395–409, Aug. 2011, publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMoa1102873. [Online]. Available: https://doi.org/10.1056/NEJMoa1102873

[14] H. J. de Koning, C. M. van der Aalst, P. A. de Jong, E. T. Scholten, K. Nackaerts, M. A. Heuvelmans, J.-W. J. Lammers, C. Weenink, U. Yousaf-Khan, N. Horeweg, S. van 't Westeinde, M. Prokop, W. P. Mali, F. A. Mohamed Hoesein, P. M. van Ooijen, J. G. Aerts, M. A. den Bakker, E. Thunnissen, J. Verschakelen, R. Vliegenthart, J. E. Walter, K. ten Haaf, H. J. Groen, and M. Oudkerk, "Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial," *New England Journal of Medicine*, vol. 382, no. 6, pp. 503–513, Feb. 2020, publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMoa1911793. [Online]. Available: https://doi.org/10.1056/NEJMoa1911793

[15] E. S. Neal Joshua, D. Bhattacharyya, M. Chakkravarthy, and Y.-C. Byun, "3D CNN with Visual Insights for Early Detection of Lung Cancer Using Gradient-Weighted Class Activation," *Journal of Healthcare Engineering*, vol. 2021, p. 6695518, 2021.

[16] Y. Li, Z. Zhang, C. Dai, Q. Dong, and S. Badrigilan, "Accuracy of deep learning for automated detection of pneumonia using chest X-Ray images: A systematic review and meta-analysis," *Computers in Biology and Medicine*, vol. 123, p. 103898, Aug. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S001048252030247X

[17] P. Lakhani and B. Sundaram, "Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks," *Radiology*, vol. 284, no. 2, pp. 574–582, Apr. 2017, publisher: Radiological Society of North America. [Online]. Available: https://pubs.rsna.org/doi/full/10.1148/radiol.2017162326

[18] H. Mohammad-Rahimi, M. Nadimi, A. Ghalyanchi-Langeroudi, M. Taheri, and S. Ghafouri-Fard, "Application of Machine Learning in Diagnosis of COVID-19 Through X-Ray and CT Images: A Scoping Review," *Frontiers in Cardiovascular Medicine*, vol. 8, 2021, publisher: Frontiers. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcvm.2021.638011/full

[19] S. M. Lee, J. B. Seo, J. Yun, Y.-H. Cho, J. Vogel-Claussen, M. L. Schiebler, W. B. Gefter, E. J. R. van Beek, J. M. Goo, K. S. Lee, H. Hatabu, J. Gee, and N. Kim, "Deep Learning Applications in Chest Radiography and Computed Tomography: Current State of the Art," *Journal of Thoracic Imaging*, vol. 34, no. 2, pp. 75–85, Mar. 2019.

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, number: 7553 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/nature14539

[21] J. van Hoek, A. Huber, A. Leichtle, K. Härmä, D. Hilt, H. von Tengg-Kobligk, J. Heverhagen, and A. Poellinger, "A survey on the future of radiology among radiologists, medical students and surgeons: Students and surgeons tend to be more skeptical about artificial intelligence and radiologists may fear that other disciplines take over," *European Journal of*

Radiology, vol. 121, p. 108742, Dec. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0720048X19303924

[22] R. L. Siddon, "Fast calculation of the exact radiological path for a three-dimensional CT array," *Medical Physics*, vol. 12, no. 2, pp. 252–255, Apr. 1985.

[23] E. Sundermann, F. Jacobs, M. Christiaens, B. De Sutter, and I. Lemahieu, "A Fast Algorithm to Calculate the Exact Radiological Path Through a Pixel Or Voxel Space," *Journal of Computing and Information Technology*, vol. 6, Dec. 1998.

[24] J. R. Van Sörnsen de Koste, H. C. J. de Boer, R. H. Schuchhard-Schipper, S. Senan, and B. J. M. Heijmen, "Procedures for high precision setup verification and correction of lung cancer patients using CT-simulation and digitally reconstructed radiographs (DRR)," *International Journal of Radiation Oncology, Biology, Physics*, vol. 55, no. 3, pp. 804–810, Mar. 2003.

[25] M. Matsushima, T. Adachi, R. Tanaka, Y. Kikuchi, M. Shimoda, T. Yoneda, and T. Yonezawa, "[Study of optimal imaging parameters for digitally reconstructed radiographs (DRR) in radiotherapy treatment planning using single-slice helical CT]," *Nihon Hoshasen Gijutsu Gakkai Zasshi*, vol. 60, no. 4, pp. 528–536, Apr. 2004.

[26] C. Yang, M. Guiney, P. Hughes, S. Leung, K. H. Liew, J. Matar, and G. Quong, "Use of digitally reconstructed radiographs in radiotherapy treatment planning and verification," *Australasian Radiology*, vol. 44, no. 4, pp. 439–443, Nov. 2000.

[27] L. Lemieux, R. Jagoe, D. R. Fish, N. D. Kitchen, and D. G. Thomas, "A patient-to-computed-tomography image registration method based on digitally reconstructed radiographs," *Medical Physics*, vol. 21, no. 11, pp. 1749–1760, Nov. 1994.

[28] C. S. Moore, G. P. Liney, A. W. Beavis, and J. R. Saunderson, "A method to produce and validate a digitally reconstructed radiograph-based computer simulation for optimisation of chest radiographs acquired with a computed radiography imaging system," *The British Journal of Radiology*, vol. 84, no. 1006, pp. 890–902, Oct. 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3473768/

[29] C. S. Moore, G. Avery, S. Balcam, L. Needler, A. Swift, A. W. Beavis, and J. R. Saunderson, "Use of a digitally reconstructed radiograph-based computer simulation for the optimisation of chest radiographic techniques for computed radiography imaging systems," *The British Journal of Radiology*, vol. 85, no. 1017, pp. e630–639, Sep. 2012.

[30] M. Unberath, J.-N. Zaech, S. C. Lee, B. Bier, J. Fotouhi, M. Armand, and N. Navab, "DeepDRR – A Catalyst for Machine Learning in Fluoroscopy-guided Procedures," *arXiv:1803.08606 [physics]*, Mar. 2018, arXiv: 1803.08606. [Online]. Available: http://arxiv.org/abs/1803.08606

[31] P. Zhang, Y. Zhong, Y. Deng, X. Tang, and X. Li, "DRR4Covid: Learning Automated COVID-19 Infection Segmentation from Digitally Reconstructed Radiographs," *arXiv:2008.11478 [cs, eess]*, Aug. 2020, arXiv: 2008.11478. [Online]. Available: http://arxiv.org/abs/2008.11478

[32] E. J. Mortani Barbosa, W. B. Gefter, F. C. Ghesu, S. Liu, B. Mailhe, A. Mansoor, S. Grbic, and S. Vogt, "Automated Detection and Quantification of COVID-19 Airspace Disease on Chest Radiographs: A Novel Approach Achieving Expert Radiologist-Level Performance Using a Deep Convolutional Neural Network Trained on Digital Reconstructed Radiographs From Computed Tomography-Derived Ground Truth," *Investigative Radiology*, Jan. 2021.

[33] M. I. Campo, J. Pascau, and R. S. José Estépar, "EMPHYSEMA QUANTIFICATION ON SIMULATED X-RAYS THROUGH DEEP LEARNING TECHNIQUES," *Proceedings. IEEE International Symposium on Biomedical Imaging*, vol. 2018, pp. 273–276, Apr. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6239425/

[34] R. M. Fuller, J. Kim, T. W. An, L. Rajan, A. D. Cororaton, P. Kumar, J. T. Deland, and S. J. Ellis, "Assessment of Flatfoot Deformity Using Digitally Reconstructed Radiographs: Reliability and Comparison to Conventional Radiographs," *Foot & Ankle International*, p. 10711007221089260, May 2022.

[35] S. Carey, S. Kandel, C. Farrell, J. Kavanagh, T. Chung, W. Hamilton, and P. Rogalla, "Comparison of conventional chest x ray with a novel projection technique for ultra-low dose CT," *Medical Physics*, vol. 48, no. 6, pp. 2809–2815, 2021, _eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.14142. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.14142

[36] B. C. Council, "Major John Hall-Edwards - Birmingham City Council," Sep. 2012. [Online]. Available: https://web.archive.org/web/20120928204852/http://www.birmingham.gov.uk/xray

[37] S. Abdullah, "Production of X-rays." [Online]. Available: https://www.radiologycafe.com/radiology-trainees/frcr-physics-notes/production-of-x-rays

[38] R. K. Hobbie and B. J. Roth, *Intermediate Physics for Medicine and Biology*, 4th ed. New York: Springer-Verlag, 2007. [Online]. Available: https://www.springer.com/gp/book/9780387309422

[39] Startpuntradiologie, "x-thorax - Startpuntradiologie.nl." [Online]. Available: https://www.startpuntradiologie.nl/coschappen/spoedeisende-hulp/thorax/x-thorax/index.html

[40] S. Abdullah, "CT equipment." [Online]. Available: https://www.radiologycafe.com/radiology-trainees/frcr-physics-notes/ct-equipment

[41] D. F. Swinehart, "The Beer-Lambert Law," *Journal of Chemical Education*, vol. 39, no. 7, p. 333, Jul. 1962, publisher: American Chemical Society. [Online]. Available: https://doi.org/10.1021/ed039p333

[42] O. Gozes and H. Greenspan, "Bone Structures Extraction and Enhancement in Chest Radiographs via CNN

Trained on Synthetic Data," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, Apr. 2020, pp. 858–861, iSSN: 1945-8452.

[43] ——, "Lung Structures Enhancement in Chest Radiographs via CT Based FCNN Training," *undefined*, 2018. [Online]. Available: /paper/Lung-Structures-Enhancement-in-Chest-Radiographs-CT-Gozes-Greenspan/12b1ecab5b9135be927c7dae31b0a34e024c17d4

[44] C. Carstens and N. Muller, "Fast Calculation of Digitally Reconstructed Radiographs using Light Fields," Nov. 2008.

[45] M. Abdellah, A. Eldeib, and M. I. Owis, "Accelerating DRR generation using Fourier slice theorem on the GPU," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2015, pp. 4238–4241, 2015.

[46] G. J. Tornai, G. Cserey, and I. Pappas, "Fast DRR generation for 2D to 3D registration on GPUs," *Medical Physics*, vol. 39, no. 8, pp. 4795–4799, Aug. 2012.

[47] M. de Greef, J. Crezee, J. C. van Eijk, R. Pool, and A. Bel, "Accelerated ray tracing for radiotherapy dose calculations on a GPU," *Medical Physics*, vol. 36, no. 9, pp. 4095–4102, Sep. 2009.

[48] H. Meyer, R. Juran, and P. Rogalla, "softMip: A Novel Projection Algorithm for Ultra-Low-Dose Computed Tomography," *Journal of computer assisted tomography*, vol. 32, pp. 480–4, May 2008.

[49] M. Abdellah, A. Abdelaziz, E. M. B. S. Eslam Ali, S. Abdelaziz, A. Sayed, M. I. Owis, and A. Eldeib, "Parallel generation of digitally reconstructed radiographs on heterogeneous multi-GPU workstations," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2016, pp. 3953–3956, Aug. 2016.

[50] M. C. Bastida-Jumilla, J. Larrey-Ruiz, R. Verdú-Monedero, J. Morales-Sánchez, and J.-L. Sancho-Gómez, "DRR and portal image registration for automatic patient positioning in radiotherapy treatment," *Journal of Digital Imaging*, vol. 24, no. 6, pp. 999–1009, Dec. 2011.

[51] S. Yoshino, K. Miki, K. Sakata, Y. Nakayama, K. Shibayama, and S. Mori, "Digital reconstructed radiography with multiple color image overlay for image-guided radiotherapy," *Journal of Radiation Research*, vol. 56, no. 3, pp. 588–593, May 2015.

[52] D. Hamano, K. Yoshida, C. Higuchi, D. Otsuki, H. Yoshikawa, and K. Sugamoto, "Evaluation of errors in measurements of infantile hip radiograph using digitally reconstructed radiograph from three-dimensional MRI," *Journal of Orthopaedics*, vol. 16, no. 3, pp. 302–306, Jun. 2019.

[53] A. Pyrros, A. Chen, J. M. Rodríguez-Fernández, S. M. Borstelmann, P. A. Cole, J. Horowitz, J. Chung, P. Nikolaidis, V. Boddipalli, N. Siddiqui, M. Willis, A. E. Flanders, and S. Koyejo, "Deep Learning-Based Digitally Reconstructed Tomography of the Chest in the Evaluation of Solitary Pulmonary Nodules: A Feasibility Study," *Academic Radiology*, pp. S1076–6332(22)00 307–5, Jun. 2022.

[54] S. Mutasa, S. Varada, A. Goel, T. T. Wong, and M. J. Rasiej, "Advanced Deep Learning Techniques Applied to Automated Femoral Neck Fracture Detection and Classification," *Journal of Digital Imaging*, vol. 33, no. 5, pp. 1209–1217, Oct. 2020. [Online]. Available: https://doi.org/10.1007/s10278-020-00364-8

[55] A. M. Winkler, "The NIFTI file format," Sep. 2012. [Online]. Available: https://brainder.org/2012/09/23/the-nifti-file-format/

[56] D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, O. Pele, and M. Werman, "The Quadratic-Chi Histogram Distance Family," in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, vol. 6312, pp. 749–762, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-642-15552-9_54

[57] T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection," *IEEE Transactions on Communication Technology*, vol. 15, no. 1, pp. 52–60, Feb. 1967, conference Name: IEEE Transactions on Communication Technology.

[58] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta math. Soc.*, vol. 35, pp. 99–109, 1943. [Online]. Available: https://cir.nii.ac.jp/crid/1572261550690788352

[59] P. Mayring, *Qualitative content analysis - theoretical foundation, basic procedures and software solution*, Jan. 2014.

[60] A. Boehm, *Chapter 5.13 "Theoretical Coding: Text Analysis in Grounded Theory"*, Jan. 2004, journal Abbreviation: A Companion to Qualitative Research Publication Title: A Companion to Qualitative Research.

[61] N. Salem, H. Malik, and A. Shams, "Medical image enhancement based on histogram algorithms," *Procedia Computer Science*, vol. 163, pp. 300–311, Jan. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050919321519

[62] M. Okada, T. Nomura, Y. Nakashima, S. Kido, and K. Ito, "Histogram-based comparison between dynamic and static lung perfused blood volume images using dual energy CT," *European Journal of Radiology*, vol. 108, pp. 269–275, Nov. 2018, publisher: Elsevier. [Online]. Available: https://www.ejradiology.com/article/S0720-048X(18)30280-8/fulltext

[63] N. Bottenus, B. Byram, and D. Hyun, "Histogram matching for visual ultrasound image comparison," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 68, no. 5, pp. 1487–1495, May 2021. [Online]. Available: https://www.ncbi.nlm.

nih.gov/pmc/articles/PMC8136614/

[64] R. H. H. M. Philipsen, P. Maduskar, L. Hogeweg, J. Melendez, C. I. Sánchez, and B. van Ginneken, "Localized Energy-Based Normalization of Medical Images: Application to Chest Radiography," *IEEE Transactions on Medical Imaging*, vol. 34, no. 9, pp. 1965–1975, Sep. 2015, conference Name: IEEE Transactions on Medical Imaging.

[65] T. Zhao, M. McNitt-Gray, and D. Ruan, "A convolutional neural network for ultra-low-dose CT denoising and emphysema screening," *Medical Physics*, vol. 46, no. 9, pp. 3941–3950, Sep. 2019.

[66] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 3462–3471, arXiv:1705.02315 [cs]. [Online]. Available: http://arxiv.org/abs/1705.02315

[67] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A Survey on Deep Learning in Medical Image Analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017, arXiv: 1702.05747. [Online]. Available: http://arxiv.org/abs/1702.05747

[68] E. Sogancioglu, E. Çallı, B. van Ginneken, K. G. van Leeuwen, and K. Murphy, "Deep Learning for Chest X-ray Analysis: A Survey," *arXiv:2103.08700 [cs, eess]*, Mar. 2021, arXiv: 2103.08700. [Online]. Available: http://arxiv.org/abs/2103.08700

[69] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," Jan. 2019. [Online]. Available: https://arxiv.org/abs/1901.07031v1

[70] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," Dec. 2017, arXiv:1711.05225 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1711.05225

[71] P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, B. N. Patel, K. W. Yeom, K. Shpanskaya, F. G. Blankenberg, J. Seekins, T. J. Amrhein, D. A. Mong, S. S. Halabi, E. J. Zucker, A. Y. Ng, and M. P. Lungren, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLoS medicine*, vol. 15, no. 11, p. e1002686, Nov. 2018.

[72] L. Yao, J. Prosky, B. Covington, and K. Lyman, "A Strong Baseline for Domain Adaptation and Generalization in Medical Imaging," Apr. 2019, arXiv:1904.01638 [cs, eess, stat]. [Online]. Available: http://arxiv.org/abs/1904.01638

[73] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLOS Medicine*, vol. 15, no. 11, p. e1002683, Nov. 2018, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002683

[74] Y.-G. Kim, Y. Cho, C.-J. Wu, S. Park, K.-H. Jung, J. B. Seo, H. J. Lee, H. J. Hwang, S. M. Lee, and N. Kim, "Short-term Reproducibility of Pulmonary Nodule and Mass Detection in Chest Radiographs: Comparison among Radiologists and Four Different Computer-Aided Detections with Convolutional Neural Net," *Scientific Reports*, vol. 9, no. 1, p. 18738, Dec. 2019, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-019-55373-7

[75] Y. Zhang, S. Miao, T. Mansi, and R. Liao, "Task Driven Generative Modeling for Unsupervised Domain Adaptation: Application to X-ray Image Segmentation," 2018, vol. 11071, pp. 599–607, arXiv:1806.07201 [cs]. [Online]. Available: http://arxiv.org/abs/1806.07201

[76] M. Lenga, H. Schulz, and A. Saalbach, "Continual Learning for Domain Adaptation in Chest X-ray Classification," Jan. 2020, arXiv:2001.05922 [cs]. [Online]. Available: http://arxiv.org/abs/2001.05922

[77] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, Jan. 2019, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41591-018-0316-z

[78] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Apr. 2015, arXiv:1409.1556 [cs]. [Online]. Available: http://arxiv.org/abs/1409.1556

[79] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385 [cs]*, Dec. 2015, arXiv: 1512.03385. [Online]. Available: http://arxiv.org/abs/1512.03385

[80] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *arXiv:1608.06993 [cs]*, Aug. 2016, arXiv: 1608.06993. [Online]. Available: http://arxiv.org/abs/1608.06993

[81] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," Jan. 2015, arXiv:1409.0575 [cs]. [Online]. Available: http://arxiv.org/abs/1409.0575

[82] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding Neural Networks Through

Deep Visualization," *arXiv:1506.06579 [cs]*, Jun. 2015, arXiv: 1506.06579. [Online]. Available: http://arxiv.org/abs/1506.06579

[83] M. Cicero, A. Bilbily, E. Colak, T. Dowdell, B. Gray, K. Perampaladas, and J. Barfett, "Training and Validating a Deep Convolutional Neural Network for Computer-Aided Detection and Classification of Abnormalities on Frontal Chest Radiographs," *Investigative Radiology*, vol. 52, no. 5, pp. 281–287, May 2017.

[84] J.-Z. Cheng, D. Ni, Y.-H. Chou, J. Qin, C.-M. Tiu, Y.-C. Chang, C.-S. Huang, D. Shen, and C.-M. Chen, "Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans," *Scientific Reports*, vol. 6, p. 24454, Apr. 2016.

[85] R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V. McConnell, G. S. Corrado, L. Peng, and D. R. Webster, "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning," *Nature Biomedical Engineering*, vol. 2, no. 3, pp. 158–164, Mar. 2018, number: 3 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41551-018-0195-0

[86] Y. Liu, K. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P. Q. Nelson, G. S. Corrado, J. D. Hipp, L. Peng, and M. C. Stumpe, "Detecting Cancer Metastases on Gigapixel Pathology Images," Mar. 2017, arXiv:1703.02442 [cs]. [Online]. Available: http://arxiv.org/abs/1703.02442

[87] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

[88] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, Dec. 2016.

[89] T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. Mann, A. den Heeten, and N. Karssemeijer, "Large scale deep learning for computer aided detection of mammographic lesions," *Medical Image Analysis*, vol. 35, pp. 303–312, Jan. 2017.

[90] A. Jamaludin, T. Kadir, and A. Zisserman, "SpineNet: Automatically Pinpointing Classification Evidence in Spinal MRIs," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds. Cham: Springer International Publishing, 2016, vol. 9901, pp. 166–175, series Title: Lecture Notes in Computer Science. [Online]. Available: https://link.springer.com/10.1007/978-3-319-46723-8_20

[91] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren, "Preparing Medical Imaging Data for Machine Learning," *Radiology*, vol. 295, no. 1, pp. 4–15, Apr. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7104701/

[92] Y.-X. Tang, Y.-B. Tang, Y. Peng, K. Yan, M. Bagheri, B. A. Redd, C. J. Brandon, Z. Lu, M. Han, J. Xiao, and R. M. Summers, "Automated abnormality classification of chest radiographs using deep convolutional neural networks," *NPJ Digital Medicine*, vol. 3, p. 70, May 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7224391/

[93] F. Grün, C. Rupprecht, N. Navab, and F. Tombari, "A Taxonomy and Library for Visualizing Learned Features in Convolutional Neural Networks," *arXiv:1606.07757 [cs]*, Jun. 2016, arXiv: 1606.07757. [Online]. Available: http://arxiv.org/abs/1606.07757

[94] K. R. Mopuri, U. Garg, and R. V. Babu, "CNN Fixations: An unraveling approach to visualize the discriminative image regions," *arXiv:1708.06670 [cs]*, Aug. 2017, arXiv: 1708.06670. [Online]. Available: http://arxiv.org/abs/1708.06670

[95] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning," *arXiv:1806.00069 [cs, stat]*, May 2018, arXiv: 1806.00069. [Online]. Available: http://arxiv.org/abs/1806.00069

[96] Q. Yang, A. Steinfeld, and J. Zimmerman, "Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes," *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–11, May 2019, arXiv: 1904.09612. [Online]. Available: http://arxiv.org/abs/1904.09612

[97] S. Guendel, S. Grbic, B. Georgescu, S. K. Zhou, L. Ritschl, A. Meier, and D. Comaniciu, "Learning to recognize Abnormalities in Chest X-Rays with Location-Aware Dense Networks," Mar. 2018.

[98] Y. Feng, H. Teh, and Y. Cai, "Deep Learning for Chest Radiology: A Review," *Current Radiology Reports*, vol. 7, Jul. 2019.

[99] B. van Ginneken, "Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning," *Radiological Physics and Technology*, vol. 10, no. 1, pp. 23–32, Mar. 2017.

[100] O. Paalvast, M. Nauta, M. Koelle, J. Geerdink, O. Vijlbrief, J. H. Hegeman, and C. Seifert, "Radiology report generation for proximal femur fractures using deep classification and language generation models," *Artificial Intelligence in Medicine*, vol. 128, p. 102281, Jun. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S093336572200046X

[101] M. M. A. Monshi, J. Poon, and V. Chung, "Deep learning in generating radiology reports: A survey,"

Artificial Intelligence in Medicine, vol. 106, p. 101878, Jun. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7227610/

[102] L. Oakden-Rayner, "Exploring Large-scale Public Medical Image Datasets," *Academic Radiology*, vol. 27, no. 1, pp. 106–112, Jan. 2020, publisher: Elsevier. [Online]. Available: https://www.academicradiology.org/article/S1076-6332(19)30488-X/fulltext

[103] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Jan. 2017, arXiv: 1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

[104] S. Patel, B. K. P, and R. K. Muthu, "Medical Image Enhancement Using Histogram Processing and Feature Extraction for Cancer Classification," Mar. 2020, arXiv:2003.06615 [cs]. [Online]. Available: http://arxiv.org/abs/2003.06615

[105] W.-Y. Hsu and C.-Y. Chou, "Medical Image Enhancement Using Modified Color Histogram Equalization," *Journal of Medical and Biological Engineering*, vol. 35, no. 5, pp. 580–584, Oct. 2015. [Online]. Available: https://doi.org/10.1007/s40846-015-0078-8

[106] S. Pizer, R. Johnston, J. Ericksen, B. Yankaskas, and K. Muller, "Contrast-limited adaptive histogram equalization: speed and effectiveness," in *[1990] Proceedings of the First Conference on Visualization in Biomedical Computing*, May 1990, pp. 337–345.

[107] D. Li, B. Mikela Vilmun, J. Frederik Carlsen, E. Albrecht-Beste, C. Ammitzbøl Lauridsen, M. Bachmann Nielsen, and K. Lindskov Hansen, "The Performance of Deep Learning Algorithms on Automatic Pulmonary Nodule Detection and Classification Tested on Different Datasets That Are Not Derived from LIDC-IDRI: A Systematic Review," *Diagnostics*, vol. 9, no. 4, p. 207, Dec. 2019, number: 4 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2075-4418/9/4/207

[108] B. Segal, D. M. Rubin, G. Rubin, and A. Pantanowitz, "Evaluating the Clinical Realism of Synthetic Chest X-Rays Generated Using Progressively Growing GANs," *SN Computer Science*, vol. 2, no. 4, p. 321, Jun. 2021. [Online]. Available: https://doi.org/10.1007/s42979-021-00720-7

[109] S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. van Ginneken, A. Madabhushi, J. L. Prince, D. Rueckert, and R. M. Summers, "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises," *Proceedings of the IEEE*, pp. 1–19, 2021, arXiv: 2008.09104. [Online]. Available: http://arxiv.org/abs/2008.09104

[110] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," Mar. 2020, arXiv:1912.04958 [cs, eess, stat]. [Online]. Available: http://arxiv.org/abs/1912.04958

[111] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," Mar. 2019, arXiv:1812.04948 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1812.04948

[112] C. Baur, S. Albarqouni, and N. Navab, "MelanoGANs: High Resolution Skin Lesion Synthesis with GANs," Apr. 2018, arXiv:1804.04338 [cs]. [Online]. Available: http://arxiv.org/abs/1804.04338

[113] J. M. Wolterink, A. M. Dinkla, M. H. F. Savenije, P. R. Seevinck, C. A. T. van den Berg, and I. Išgum, "Deep MR to CT Synthesis Using Unpaired Data," in *Simulation and Synthesis in Medical Imaging*, ser. Lecture Notes in Computer Science, S. A. Tsaftaris, A. Gooya, A. F. Frangi, and J. L. Prince, Eds. Cham: Springer International Publishing, 2017, pp. 14–23.

[114] X. Yi, E. Walia, and P. Babyn, "Generative Adversarial Network in Medical Imaging: A Review," *Medical Image Analysis*, vol. 58, p. 101552, Dec. 2019, arXiv: 1809.07294. [Online]. Available: http://arxiv.org/abs/1809.07294

[115] A. Beers, J. Brown, K. Chang, J. P. Campbell, S. Ostmo, M. F. Chiang, and J. Kalpathy-Cramer, "High-resolution medical image synthesis using progressively grown generative adversarial networks," May 2018. [Online]. Available: https://arxiv.org/abs/1805.03144v2

[116] H. Salehinejad, E. Colak, T. Dowdell, J. Barfett, and S. Valaee, "Synthesizing Chest X-Ray Pathology for Training Deep Convolutional Neural Networks," *IEEE transactions on medical imaging*, vol. 38, no. 5, pp. 1197–1206, May 2019.

[117] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," Jun. 2014, arXiv:1406.2661 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1406.2661

[118] P. Zhang, F. Wang, W. Xu, and Y. Li, "Multi-channel Generative Adversarial Network for Parallel Magnetic Resonance Image Reconstruction in K-space," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, ser. Lecture Notes in Computer Science, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham: Springer International Publishing, 2018, pp. 180–188.

[119] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "Generative Adversarial Networks for Noise Reduction in Low-Dose CT," *IEEE Transactions on Medical Imaging*, vol. 36, no. 12, pp. 2536–2545, Dec. 2017, conference Name: IEEE Transactions on Medical Imaging.

[120] J. Son, S. J. Park, and K.-H. Jung, "Retinal Vessel Segmentation in Fundoscopic Images with Generative Adversarial Networks," Jun. 2017, arXiv:1706.09318 [cs]. [Online]. Available: http://arxiv.org/abs/1706.09318

[121] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic Data Augmentation using GAN for Improved Liver Lesion Classification," Jan. 2018, arXiv:1801.02385 [cs]. [Online]. Available: http://arxiv.org/abs/1801.02385

[122] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," Jan. 2016, arXiv:1511.06434 [cs]. [Online]. Available: http://arxiv.org/abs/1511.06434

[123] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," Feb. 2018, arXiv:1710.10196 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1710.10196

[124] L. Mescheder, A. Geiger, and S. Nowozin, "Which Training Methods for GANs do actually Converge?" Jul. 2018, arXiv:1801.04406 [cs]. [Online]. Available: http://arxiv.org/abs/1801.04406

[125] S. Sundaram and N. Hulkund, "GAN-based Data Augmentation for Chest X-ray Classification," Jul. 2021, arXiv:2107.02970 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2107.02970

[126] M. Moradi, A. Madani, A. Karargyris, and T. Syeda-Mahmood, "Chest x-ray generation and data augmentation for cardiovascular abnormality classification," Mar. 2018, p. 57.

[127] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 6, pp. 1153–1160, Dec. 1981, conference Name: IEEE Transactions on Acoustics, Speech, and Signal Processing.

[128] K. Shmelkov, C. Schmid, and K. Alahari, "How good is my GAN?" Jul. 2018, arXiv:1807.09499 [cs]. [Online]. Available: http://arxiv.org/abs/1807.09499

[129] A. Brock, J. Donahue, and K. Simonyan, "Large Scale GAN Training for High Fidelity Natural Image Synthesis," Feb. 2019, arXiv:1809.11096 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1809.11096

[130] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. C. Loy, Y. Qiao, and X. Tang, "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks," Sep. 2018, arXiv:1809.00219 [cs]. [Online]. Available: http://arxiv.org/abs/1809.00219

[131] C.-Y. Zhao, R.-S. Jia, Q.-M. Liu, X.-Y. Liu, H.-M. Sun, and X.-L. Zhang, "Chest X-ray images super-resolution reconstruction via recursive neural network," *Multimedia Tools and Applications*, vol. 80, no. 1, pp. 263–277, Jan. 2021. [Online]. Available: https://doi.org/10.1007/s11042-020-09773-x

[132] S. Dai, M. Han, W. Xu, Y. Wu, and Y. Gong, "Soft Edge Smoothness Prior for Alpha Channel Super Resolution," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2007, pp. 1–8, iSSN: 1063-6919.

[133] Q. Shan, "Fast image/video upsampling | ACM Transactions on Graphics." [Online]. Available: https://dl.acm.org/doi/abs/10.1145/1409060.1409106

[134] W. Freeman, T. Jones, and E. Pasztor, "Example-based super-resolution," *IEEE Computer Graphics and Applications*, vol. 22, no. 2, pp. 56–65, Mar. 2002, conference Name: IEEE Computer Graphics and Applications.

[135] C. Dong, C. C. Loy, K. He, and X. Tang, "Image Super-Resolution Using Deep Convolutional Networks," Jul. 2015, arXiv:1501.00092 [cs]. [Online]. Available: http://arxiv.org/abs/1501.00092

[136] J. Kim, J. K. Lee, and K. M. Lee, "Accurate Image Super-Resolution Using Very Deep Convolutional Networks," Nov. 2016, arXiv:1511.04587 [cs]. [Online]. Available: http://arxiv.org/abs/1511.04587

[137] Y. Tai, J. Yang, and X. Liu, "Image Super-Resolution via Deep Recursive Residual Network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 2790–2798, iSSN: 1063-6919.

[138] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," Mar. 2016, arXiv:1603.08155 [cs]. [Online]. Available: http://arxiv.org/abs/1603.08155

[139] J. Bruna, P. Sprechmann, and Y. LeCun, "Super-Resolution with Deep Convolutional Sufficient Statistics," Mar. 2016, arXiv:1511.05666 [cs]. [Online]. Available: http://arxiv.org/abs/1511.05666

[140] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," May 2017, arXiv:1609.04802 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1609.04802

[141] B. C. Maral, "Single Image Super-Resolution Methods: A Survey," Feb. 2022, arXiv:2202.11763 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2202.11763

[142] Y. Li, B. Sixou, and F. Peyrin, "A Review of the Deep Learning Methods for Medical Images Super Resolution Problems," *IRBM*, vol. 42, no. 2, pp. 120–133, Apr. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1959031820301408

[143] K. Zhang, X. Gao, D. Tao, and X. Li, "Single Image Super-Resolution With Non-Local Means and Steering Kernel Regression," *IEEE Transactions on Image Processing*, vol. 21, no. 11, pp. 4544–4556, Nov. 2012, conference Name: IEEE Transactions on Image Processing.

[144] K. Prajapati, V. Chudasama, H. Patel, K. Upla, R. Ramachandra, K. Raja, and C. Busch, "Unsupervised Single Image Super-Resolution Network (USISResNet) for Real-World Data Using Generative Adversarial Network," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2020, pp. 1904–1913, iSSN: 2160-7516.

[145] A. Lugmayr, M. Danelljan, and R. Timofte, "Unsupervised Learning for Real-World Super-Resolution," Sep. 2019, arXiv:1909.09629 [cs, eess]. [Online]. Available: http://arxiv.org/abs/1909.09629

[146] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced Deep Residual Networks for Single Image Super-Resolution," Jul. 2017, arXiv:1707.02921 [cs]. [Online]. Available: http://arxiv.org/abs/1707.02921

[147] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual Dense Network for Image Super-Resolution," Mar. 2018, arXiv:1802.08797 [cs]. [Online]. Available: http://arxiv.org/abs/1802.08797

[148] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, Accurate, and Lightweight Super-Resolution with Cascading Residual Network," Oct. 2018, arXiv:1803.08664 [cs]. [Online]. Available: http://arxiv.org/abs/1803.08664

[149] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 PIRM Challenge on Perceptual Image Super-resolution," Jan. 2019, arXiv:1809.07517 [cs]. [Online]. Available: http://arxiv.org/abs/1809.07517

[150] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004, conference Name: IEEE Transactions on Image Processing.

[151] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, Jan. 2009, conference Name: IEEE Signal Processing Magazine.

[152] G. M. S and B. V. R, "Super-resolution using Deep Networks for Chest X-Ray Images," in *2021 Sixth International Conference on Image Information Processing (ICIIP)*, vol. 6, Nov. 2021, pp. 198–201, iSSN: 2640-074X.

[153] J. Park, D. Hwang, K. Y. Kim, S. K. Kang, Y. K. Kim, and J. S. Lee, "Computed tomography super-resolution using deep convolutional neural network," *Physics in Medicine &amp$\mathsemicolon$ Biology*, vol. 63, no. 14, p. 145011, Jul. 2018, publisher: IOP Publishing. [Online]. Available: https://doi.org/10.1088/1361-6560/aacdd4

[154] K. Zeng, H. Zheng, C. Cai, Y. Yang, K. Zhang, and Z. Chen, "Simultaneous single- and multi-contrast super-resolution for brain MRI images based on a convolutional neural network," *Computers in Biology and Medicine*, vol. 99, pp. 133–141, Aug. 2018.

[155] L. Xu, X. Zeng, Z. Huang, W. Li, and H. Zhang, "Low-dose chest X-ray image super-resolution using generative adversarial nets with spectral normalization," *Biomedical Signal Processing and Control*, vol. 55, p. 101600, Jan. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809419301818

[156] J. Zhang, C. Xu, J. Li, Y. Han, Y. Wang, Y. Tai, and Y. Liu, "SCSNet: An Efficient Paradigm for Learning Simultaneously Image Colorization and Super-Resolution," Jan. 2022, arXiv:2201.04364 [cs]. [Online]. Available: http://arxiv.org/abs/2201.04364

[157] W. Xing and K. Egiazarian, "End-to-End Learning for Joint Image Demosaicing, Denoising and Super-Resolution," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 3506–3515, iSSN: 2575-7075.

[158] Z. Li, S. Zhang, J. Zhang, K. Huang, Y. Wang, and Y. Yu, "MVP-Net: Multi-view FPN with Position-Aware Attention for Deep Universal Lesion Detection," Oct. 2019, pp. 13–21.