A Class-Conditioned Deep Neural Network to Reconstruct CT Volumes from X-Ray Images: Depth-Aware Connection and Adaptive Feature Fusion

By

Md. Aminur Rab Ratul

A thesis submitted in partial fulfillment of the requirements for the Master's degree in Computer Science Concentration in Applied Artificial Intelligence

Ottawa-Carleton Institute for Electrical and Computer Engineering

School of Electrical Engineering and Computer Science

University of Ottawa

January 2021

© Md. Aminur Rab Ratul, Ottawa, Canada, 2021

Author’s Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis. I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices.

Statement of Contributions

The body of this thesis is based on an accepted work to IEEE International Symposium on Biomedical Imaging 2021 (ISBI 2021):

Ratul, A. R., Yuan, K., & Lee, W. S. (2020). CCX-rayNet: A Class Conditioned Convolutional Neural Network for Biplanar X-rays to CT Volume. (Accepted to ISBI2021)

I would like to acknowledge my lab member Kun Yuan who contributed to the research described in this dissertation.

Chapter 1: We find the shortcomings of the present reconstruction model and probable solutions. Specifically, he helped me to write section 1.3 (Motivation and Challenges), 1.4 (Objective), and 1.5 (Contributions of This Work)

Chapter 3: Kun Yuan and I discussed the Methodology together and proposed three new modules (DFT, DAC, AFF). Furthermore, he also assists me in setting up the training process of two networks (CCX-rayNet and CCX-rayGAN).

Chapter 5: After achieving the results from our networks, we critically discuss the outcome and try to find the efficacy of our proposed methods.

Abstract

X-ray and Computed Tomography (CT) are two of the most usual imaging systems for the medical domain in order to inspect the body's inside, check the major organs, and internal injuries. While the Computed Tomography (CT) generates 3D images to check internal body structure for diagnosis, the X-ray has offered a cost-effective imaging technique and incurs lower radiation dose. Therefore, due to the high screening cost and deficiency of screening machines, CT scan appliances are not always accessible. 3D CT volume reconstruction from its 2D X-ray counterpart is a challenging problem. Furthermore, the conventional CT reconstruction algorithms necessitate hundreds of X-ray projections to provide the body's 3D view that cannot be executed on a traditional X-ray apparatus. In spite of the advancement of convolutional neural networks to achieve a reconstruction performance by generating the anatomical structures with realistic textures, semantic features still remain unexplored. In this study, we investigate the novel 3D reconstruction problem by proposing a class-conditioned deep neural network (CCX-rayNet), which is adept at recovering textures and shapes faithful to semantic classes in the real-world. Firstly, we propose to modulate the features conditioned on the categorical prior by generating the affine transformation parameters. Secondly, we enhance the feature representation of the X-ray image by bridging 2D and 3D features. In particular, we estimate a 3D attention mask to be applied on the expanded 3D feature map, where the contextual relationship will be highlighted. In our proposed biplanar view CCX-rayNet, we also include an adaptive feature fusion (AFF) module to resolve a registration problem that happens with uncontrolled input data by using the similarity matrix. To the best of our knowledge, this proposed method is the first work to deal with semantic prior in the 3D CT reconstruction task. Both qualitative and quantitative assessments demonstrate the exceptional performance of our proposed CCX-rayNet, which outperforms state-of-the-art methods. Using this 3D reconstruction process, a medical practitioner can achieve some useful appliances for clinical practices at a minimal cost.

Publication During Master's

Peer-Reviewed

1. "PS8-Net: A Deep Convolutional Neural Network to Predict the Eight-State Protein Secondary Structure", Md Aminur Rab Ratul, Maryam Tavakol Elahi, M. Hamed Mozaffari, WonSook Lee, accepted as a short paper with poster to The International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2020
2. "RAM-Net: A Residual Attention MobileNet toDetect COVID-19 Cases from Chest X-Ray Images", Md Aminur Rab Ratul, Maryam Tavakol Elahi, Kun Yuan, WonSook Lee, accepted as a short paper with oral presentation to IEEE International Conference on Machine Learning and Applications (IEEE ICMLA) 2020
3. "CCX-rayNet: A Class Conditioned Convolutional Neural Network for Biplanar X-rays to CT Volume", Md Aminur Rab Ratul, Kun Yuan, WonSook Lee, accepted to IEEE International Symposium on Biomedical Imaging 2021 (ISBI 2021).
4. "IrisNet: Deep Learning for Automatic and Real-time Tongue Contour Tracking in Ultrasound Video Data using Peripheral Vision", M. Hamed Mozaffari, Md Aminur Rab Ratul, WonSook Lee, submitted to international journal of pattern recognition and artificial intelligence (https://arxiv.org/ftp/arxiv/papers/1911/1911.03972.pdf)

Not Peer-Reviewed

1. "Prediction of 8-state protein secondary structures by 1D-Inception and BDLSTM", Md Aminur Rab Ratul, Marcel Turcotte, M. Hamed Mozaffari, WonSook Lee. (https://www.biorxiv.org/content/biorxiv/early/2019/12/12/871921.full.pdf)
2. "Skin Lesions Classification Using Deep Learning Based on Dilated Convolution", Md Aminur Rab Ratul, M. Hamed Mozaffari, WonSook Lee, Enea Parimbelli. (https://www.biorxiv.org/content/biorxiv/early/2020/06/05/860700.full.pdf)
3. "Atrous Convolution with Transfer Learning for Skin Lesions Classification", Md Aminur Rab Ratul, M. Hamed Mozaffari, WonSook Lee, Enea Parimbelli. (https://www.biorxiv.org/content/biorxiv/early/2019/09/14/746388.full.pdf)
4. "Inspect Transfer Learning Architecture with Dilated Convolution", Syeda Noor Jaha Azim, Md Aminur Rab Ratul. (https://arxiv.org/ftp/arxiv/papers/1911/1911.08769.pdf)

Dedication

I would like to dedicate this thesis to my mother, Nura Nazmin. Her sacrifices, encouragement, direction, and confidence helped me evolve and inspired me to accomplish good deeds.

Acknowledgment

First and foremost, I would like to express my deepest gratitude to my supervisor Professor WonSook Lee for giving me unlimited support and inspiration over the past two years. She always helps me to think intuitively to pursue any research questions and encourages me to keep on track to achieve my dream. I can not be able to accomplish this degree without her proper guidance and dedicated involvement in every step of my research work.

In the last two years, I have come and contact with some incredible and helpful lab members (Hamed, Maryam, Qing, Kun, Chanho, Iffa, Kang, Jiawei, just to name a few). They assist me throughout my time at the University of Ottawa with constructive feedback, support, and in-depth knowledge.

Last but not least, I would like to acknowledge the tremendous love, motivation, and immense support my family and friends had given me. Specifically, without the inspiration from my parents, I would not be able to reach here.

Contents

List of Figures

List of Tables

List of Equation

Epigraph

"By the time we get to the 2040s, we'll be able to multiply human intelligence a billion-fold. That will be a profound change that's singular in nature. Computers are going to keep getting smaller and smaller. Ultimately, they will go inside our bodies and brains, and make us healthier, make us smarter."

- Ray Kurzweil (American writer, inventor, and futurist)

Abbreviations

AFF: Adaptive Feature Fusion

CBCT: Cone Beam Computed Tomography

CNN: Convolutional Neural Network

CPU: Central Processing Unit

CT: Computed Tomography

CXR: Chest X-ray

DAC: Depth-Aware Connection

DCNN: Deep Convolutional Neural Network

DFT: Deep Feature Transform

DNN: Deep Neural Network

DRR: Digitally Reconstructed Radiograph

FCN: Fully Convolutional Network

GAN: Generative Adversarial Network

GPU: Graphics Processing Unit

LS-GAN loss: Least Square GAN Loss

MRI: Magnetic Resonance Imaging

MSE: Mean Squared Error

PA View: Posterior-Anterior View

PSNR: Peak Signal-to-Noise Ratio

ReLU: Rectified Linear Unit

RMsProp: Root Mean Square Propagation

SGD: Stochastic Gradient Descent

SSIM: Structural Similarity Index Measure

Chapter 1 **Introduction**

Medical imaging techniques are upgrading the diagnosis and treatment of various medical conditions both for adults and children. Many medical imaging strategies exist, such as fluoroscopy, Computed tomography (CT), and radiography (conventional X-ray including mammography). These imaging technologies exert ionizing radiation to originate images of the body. X-ray and Computed Tomography (CT) are two different medical imaging techniques where physicians can inspect significant organs, blood vessels, and tissues in 2D or 3D projection. On the one hand, it is vital to use CT or 3D projection to check the internal anatomies clearly, but on the other hand, CT releases an excessive amount of radiation (which can cause cancer) and less accessible in developing countries. Therefore, it is fundamental to create a method to reconstruct CT volume from single or few X-ray images, but this still achieves less attention. The advancement of deep learning in computer vision inspired us to develop a 3D reconstruction technique called CCX-rayNet to provide CT volume from at most two X-ray images.

In the first two sections of this thesis's introductory part, we provide a detailed explanation of X-ray and CT imaging techniques and the pros and cons of these two procedures. Then, we describe the motivation, challenges, and objectives of this research work. Finally, we depict the significant contributions of this work.

1.1 **X-ray**

German scientist Wilhelm Röntgen discovered X-ray or X-radiation on November 8th, 1987, which contains high electromagnetic radiation. Instantly, after X-ray's invention, it has been extensively utilized in a wide range of medical applications. On January 11th, 1896, John Hall-Edwards from Birmingham, England, first practically applied X-ray images for clinical issues and surgical operation. It is the first radiology imaging technique that allows us to see through anyone's skin and the bones underneath it, to give us the idea of internal anatomies during the diagnosis. The X-ray radiograph images manifest the internal organs and bones in different white and black shades because other tissues assimilate a specific amount of radiation.

1.1.1 The problem of X-ray Machine

It is sometimes impossible to detect muscle damage, internal organs, and soft tissues through X-ray imaging because of its 2D projection, also known as conventional radiography or projectional radiography. Some of the popular projectional radiography techniques to show the organs are mammography, chest radiography, orbital radiography, dental radiography, etc. In this 2D projection of an X-ray image, bones and the organs' shape can be clearly visible, but all soft tissues overlap each other and are difficult to perceive. Most of the time, it is also challenging to discern the damages of the internal organs through the 2D projection of images.

1.2 Computed Tomography (CT)

The Computed Tomography (CT) scanning technique rotates on an axis and takes several 2D images (a set of X-Ray images) of a person's body from multiple angles around the body. In the CT scan, the machine's generated signal is later processed by its computer to produce cross-sectional 2D images or "slices" of the body. Then, the computer will set all the cross-sectional images on its screen and produce a 3D projection of the body's inside to exhibit the presence of injury or diseases. In a CT scan, all the tissues are displayed in 3D space; thus, it resolves X-ray images overlying problem. Moreover, it may help physicians and doctors in several other ways:

- Computed Tomography yield detailed images of soft tissue, bones, blood vessels, and internal organs.
- CT scan can be utilized to diagnose musculoskeletal disorders, heart disease, trauma, lung nodules, liver masses, cancer, appendicitis, and many more infectious diseases.
- Detect the precise location of infections, blood clot, and a tumor.
- Identify internal bleeding and injuries.
- Conduct many critical medical procedures like radiation therapy, biopsy, and surgery.

1.2.1 **The Disadvantage of CT Scan**

A CT scan is a combination of many X-ray images that use ionizing radiation, so a patient can easily be exposed to it. A CT scan exposes individuals to more radiation than any other medical imaging test such as mammograms and X-ray radiographic images because of the detailed results. In addition, these radiations can damage the cell's DNA and enhance the possibility that they will turn cancerous. Thus, numerous CT scan over time can inflate the chance of cancer. Additionally, cancer risk is enhanced in children receiving CT scans, specifically to the chest and abdomen.

The most common risk in CT scans is an allergic reaction to intravenous contrast material, especially iodine-based liquid given in the vein to visible many organs and structures on the CT scans. Some usual symptoms are itching, hives, rash, or temperature increase throughout the body. However, a severe allergic reaction to intravenous contrast, namely anaphylactic reaction, causes extreme difficulty in breathing and painful hives. These two reactions are relatively rare and can be life-threatening if not treated timely. Furthermore, kidney failure is another extremely infrequent complication of the intravenous contrast material utilized in Computed Tomography. Dehydrated people, people who have diabetes, and patients who already have kidney problems are most vulnerable to this reaction.

During pregnancy, the CT scan can be harmful to the unborn baby because of radiation exposure. Specifically, during the first trimester, it may cause a possible risk for the fetus. Therefore, we have to encourage medical officials to use the slightest amount of radiation when executing any medical imaging exams. Besides, the CT scan machine is less accessible in developing countries because it is costlier than the X-ray machine. Additionally, it creates a substantial financial burden for patients, especially those who need multiple CT examinations [1].

The CT scanning machine is less accessible not only in the developing countries but also in some parts of the developed countries. On the other hand, X-ray is available in almost all the hospitals and the diagnostics center. In addition, the CT scan is more expensive than the X-ray and the waiting time for the result is usually long.

1.3 Motivation and Challenges

Recently, deep learning methods have been extensively explored due to their strong capability in image reconstruction. In [2], Ying et al. adopt orthogonal views of X-rays to learn the well-constrained human body anatomy using the generative adversarial network (GAN) [3]. Though it proposes three kinds of connections to bridge 2D and 3D feature maps, it still suffers from the lost depth information in the 2D perspective projection X-ray feature map. Moreover, their method imposes a strict restriction on the patient's movement during the X-ray capturing procedure, leaving the registration problem behind. This unsolved registration problem induces severe artifacts when the patient moves a lot.

The deep neural network is so successful these days in the image reconstruction task. However, in the medical domain, we need a more precise method to achieve all the necessary anatomical details. Specifically, in the Chest CT volume, it is more difficult to reconstructed all the complex shapes, blood vessels, veins, important parts like hurt, lungs, etc. CT reconstruction tasks are suffering from the following problems with a deep learning-based method.

1. **Shape distortion**: GAN is widely applied to obtain perceptually feasible textures. However, this imposes the structural distortion to the shape of body anatomies (e.g., lung, heart, and liver), as shown in Figure 1-2e. Besides, conventional GAN has a weak constraint to the image generation process, leading to ambiguous textures (e.g., drawing internal lung textures to the heart area).
2. **Loss of depth information:** In the UNet-like 2D to 3D model, connection bridges 2D and 3D feature maps by expanding the 2D feature map multiple times. However, the depth information in the perspective projection is lost because each slice of feature map in 3D space is a subset of the 2D feature map. This loss of information makes the conventional expanding operation depth-unaware and hard to converge.

3. **Registration problem:** It is easy to extend the model from a single view version into a multiple view version, i.e., biplanar views, by fusing the orthogonal 3D feature maps. However, there remains a barrier to be solved towards the practical scenario, the registration problem. In the previous method, the average sum operation is adopted to fuse the features from orthogonal views. That does not consider the movement of the patient during the biplanar X-ray images capturing, leading the feature offset that makes the feature fusion unfeasible to the realistic scenario.

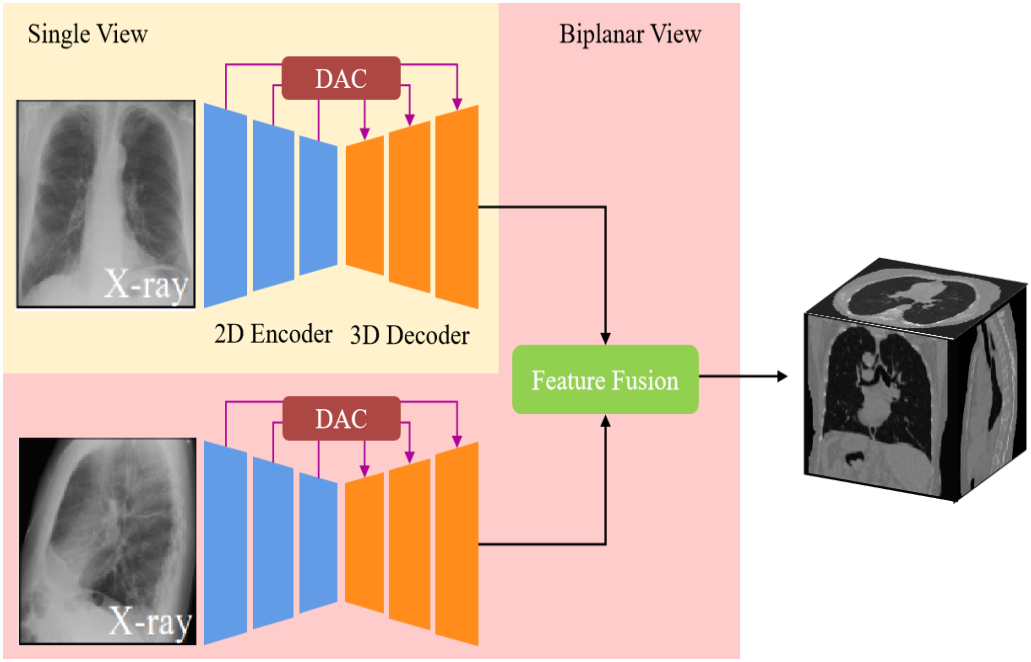

Figure 1-1. Overview of our proposed biplanar CCX-rayNet.

1.4 Overview of Proposed Methods

In this work, CCX-rayNet achieves 3D reconstruction facilitated by few X-ray images and their categorical prior. Unlike the previous methods, hundreds of X-rays are required to reconstruct a CT volume accurately [4]; our approach follows the X2CT-GAN [2] to take at most two X-ray images as input, as shown in Figure 1-1. Besides, we propose an adaptive feature fusion module to address the registration problem that happened at the rotation process, which is unaddressed in

the previous work towards the real scenario. This allows the hospital a flexible choice in selecting the X-ray apparatus. A standard X-ray machine that can do the rotation can replace the expensive one with fast rotation capability [2]. Our proposed CCX-rayNet consists of three novel modules, deep feature transform (DFT) module, depth-aware connection (DAC), and adaptive feature fusion (AFF) module, to address the challenging problems above, respectively.

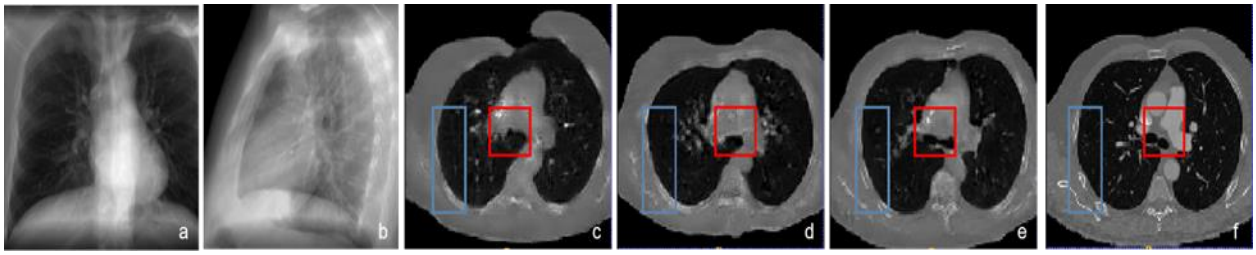

Figure 1-2. An example of CT reconstruction. (a) (b): Input X-rays. (c) X2CT result from (a) and (b). (d) X2CT result from (a) only. (e) Our result from (a) and (b). (f) Ground truth.

In particular, we show that categorical prior is crucial for reconstructing organs with different semantics (e.g., lung, liver, rib, and spine). In CT volume, different internal anatomical structures have distinct textures and should be treated differently. Inspired by [5], we modulate the 2D feature maps conditioned on the semantic information spatially in the DFT module. In this way, we can generate realistic textures faithful to the inherent class and preserve the shape concomitantly, which addresses Challenge #1 implicitly. To address Challenge #2, Depth-aware connection (DAC) bridges the encoded 2D feature and decoded 3D feature in a depth-aware manner. Consistent with perspective projection from CT volume to X-ray image, the 3D feature should be inversely projected into its corresponding encoded 2D feature. Hence, we first expand the 2D feature map into 3D and then apply attention maps to the 3D feature map slice by slice, which will help the network understand the depth information. To generalize the reconstruction model from the single view into a biplanar view, we need to train two branches of models and then fuse their features. Here, we replace the average sum operation with the weighted sum operation to bypass the less contributed feature. The weight matrix is based on a similarity matrix, and it is optimized end to end. It provides a feasible and efficient solution to Challenge #3.

1.5 Objective

This study investigates the novel 3D reconstruction problem by proposing a class-conditioned deep neural network (CCX-rayNet), which is adept at recovering textures and shapes faithful to semantic classes in the real-world. At first, we try to reduce shape distortion through the depth feature transform (DFT) module. Next, utilizing a Depth-aware connection (DAC) module to prevent depth information loss. Finally, the Adaptive feature fusion (AFF) module present to solve the registration problem from the network.

1.6 Contributions of This Work

The contributions of this work are three folds:

1. We are the first to address the 2D X-ray image to 3D CT reconstruction, given the semantic prior constraint to the extent of our knowledge. A novel CCX-rayNet is proposed to generate local visual patterns to certain internal anatomical structures adaptively. The proposed Deep Feature Transform (DFT) module achieves it in a parameter-efficient manner.
2. Depth-aware connection (DAC) is proposed to compensate for the information loss when bridging 2D, and 3D features, making the link more realistic and reducing abundant features.
3. We propose the Adaptive Feature Fusion (AFF) module to fuse multiple views of features. It suppresses the unregistered features and amplifies the most effective ones by attention mechanism.

1.7 Structure of the Thesis

The rest of the thesis is arranged into the following chapters:

- Chapter 2 provides a brief literature review, including cross-modality transformation, 3D reconstructions from 2D projections, computed tomography reconstruction from X-ray images.
- Chapter 3 exhibits the background study of deep neural networks, parameters of convolutional neural networks, and the image reconstruction process, which we will be employing in our proposed methods.
- Chapter 4 presents our proposed single view and biplanar view models, the description of the proposed modules, the objective function for the methods, and the explanation of these architecture's parameters.
- Chapter 5 explains our produced dataset for training the proposed architectures in this project. A detailed explanation of the making procedure of semantic segmentation maps has been provided.
- Chapter 6 elaborates on the training settings, optimization function, evaluation studies, qualitative analysis, quantitative analysis, and discussion.
- Chapter 7 concludes the thesis work. It also outlines the significant contributions, potential applications, and the directions of the future work.

Chapter 2 **Literature Review**

In classic direct volume rendering problem from the 3D volume, 2D rendering can be generated [6]. Drebin et al. [6] produce a system for rendering images Of volumes holding mixtures of materials. The shading model grants both the interior and boundary between all material to be colored. This work aims to give a solution for the opposite problem, which is generating 3D volume from 2D synthetic X-ray data. In Computed Tomography (CT), the volume is reconstructed from a multitude of angles or multiple views [7]. Furthermore, the quality of tomographic reconstruction has been enhanced in the past by using some popular principles like sparsity [8] or maximum likelihood [9].

We organize this section into three-part: cross-modality transformation, 3D image reconstructions from 2D image projections, computed tomography reconstruction from the X-ray image. In the cross-modality transformation part, we displayed several works that focus on image reconstruction from one-dimensionality to another. We exhibit some popular methods to reconstruct 3D images from 2D in the second part. In the final section, we mainly manifest X-ray to computed tomography (CT) reconstruction.

2.1 **Cross-Modality Transformation**

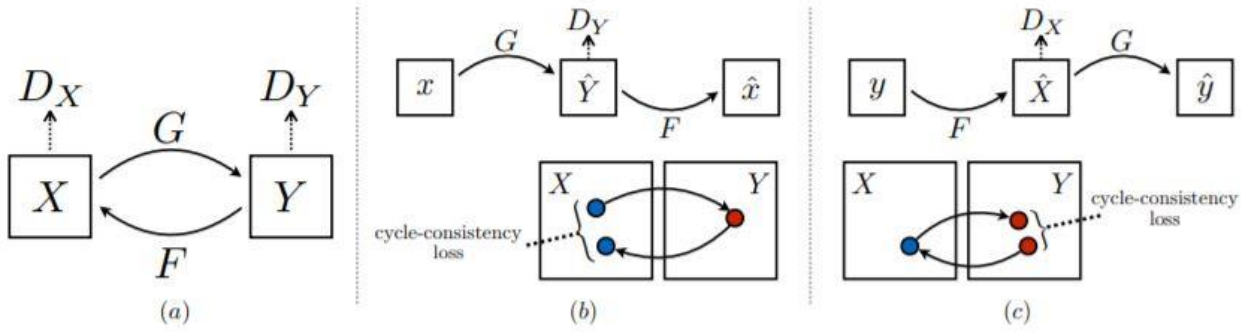

Figure 2-1. (a) two mapping function of cycle-GAN: G:X→Y and F: Y→X, and D_X and D_Y are associated adversarial discriminators, (b) forward cycle-consistent loss of cycle-GAN: x→G(x)→F(G(x))≈x, and (c) backward cycle-consistent loss of cycle-GAN: y→F(y)→G(F(y))≈y, reprinted from [13].

A deep learning architecture often required an appropriate amount of data to overcome the tendency to fall into suboptimal points during the training phase or get rid of overfitting. Synthetic data has been utilized during the training period to relieve this issue [10][11]. Provide realistic synthesized images close adjacent to target distribution is a demanding aspect of this area. Previously, Isola et al. [12] yield the pix2pix method to do the image to image transfer at the pixel level with conditional adversarial networks. CycleGAN [13], the cycle-consistent adversarial network retains the mapping between the unpaired image to image transformation. Previously there are several architectures produce in the medical image community to transfer source modality to target modalities such as MRI to CT [14][15], 3T MRI to 7T MRI [16], however where input and output were consistent (2D to 2D or 3D to 3D). The main focus of cross-modality transformation in medical images is to yield the non-linear relationship of different modalities. Here, we wish to transfer 2D chest X-ray images to a 3D CT volume. Our main aim is to provide a high-quality 3D CT volume from both the single and biplanar X-ray images with clinical values.

2.2 3D Image Reconstruction from 2D Image Projections

3D reconstruction from 2D image projections is a well-established subject in computer vision [17][18][19]. According to views represented by the 2D image, two types of 3D image reconstruction model can be provided from the 2D image: single-view reconstruction and multi-view reconstruction. Firstly, a single-view reconstruction usually requires several supplementary information to achieve the knowledge of object shape and details. As such, Yang et al. [20] proposed a 3D object reconstruction unified framework based on a small number of pose-annotated images. Wang et al. [21] present a deep implicit surface network, or DISN, to produce a high-quality detail-rich 3D mesh from a two-dimensional input image by anticipating the underlying signed distance fields. The architecture combined both the local and global features to enhance the quality of resultant 3D volume. ShapeHD [22] combines prior knowledge of adversary learned shapes with a deep generating model to pushing the bound of single-view shape reconstruction and completion. Here, prior knowledge serves as a regularizer to penalizes the network only if the outcome is impractical. Fernandez et al. [23] proposed a three-dimensional layout recovery system of the indoor scenarios from a 360° panoramic picture. This method integrates strategical

accuracy, which is achieved by geometric reasoning with the top-level pattern recognition and data abstraction process attained by deep learning.

Table 1. Several popular methods to reconstruct the 3D image from a different view of 2D image input.

Research Work	Method	Single View	Multiple View	Output Format
Yang et al. [20]	Deep convolutional neural network (DCNN)	✓	×	3D voxel
DISN [21]	Deep implicit surface network	✓	×	3D mesh
ShapeHD [22]	Generative Adversarial Network (GAN)	✓	×	3D voxel
Fernandez et al. [23]	Deep Neural Network with geometric reasoning	✓	×	3D layout
Caponetti et al. [62]	polygon mesh, B-spline interpolation, and back-lighting projections	×	✓	3D voxel
Kolev et al. [24]	Probabilistic formulation of 3D reconstruction and joint silhouette extraction	×	✓	3D reconstruction and joint silhouette extraction
3D-R2N2 [25]	3D Recurrent Reconstruction Neural Network	×	✓	3D occupancy grid
Pix2Vox [26]	Encoder-decoder based architecture	✓	✓	3D volume

On the other hand, multi-view reconstruction methods can provide the object's shape information from different angles; thus, a little prior knowledge is required. For instance, Back in 1990, Caponetti et al. [62] designed a bone reconstructed method from two X-ray image inputs based on projections of the polygon mesh, B-spline interpolation, and back-lighting projections. Kolev et al. [24] enumerate the probabilistic formulation of 3D reconstruction and joint silhouette extraction that gives rise to the observed color information from a series of calibrated two-dimensional images. 3D Recurrent Reconstruction Neural Network or 3D-R2N2 [25] retain shape advances to gain robust 3D reconstructions. This 3D reconstruction network takes one or multiple synthetic images of a particular object from random viewpoints and yields a 3D occupancy grid to show the

object's reconstructions. Besides, some research project provides three-dimensional reconstruction from both single view and multiple views of a 2D image. For example, Pix2Vox [26] is a context-aware 3D reconstruction model from both the single view and multiple views of input images. Pix2Vox is a well-organized encoder-decoder based architecture that yields a coarse 3D volume from each 2D image.

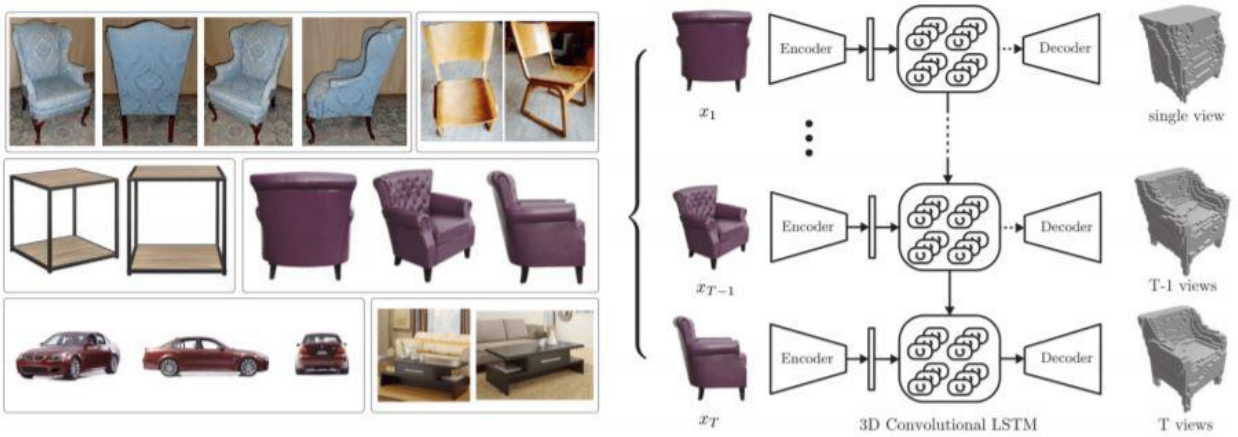

Figure 2-2. A basic structure of the 3D-R2N2 network. (a) images of different objects they want to reconstruct (left), (b) Overview of this architecture (right), reprinted from [25].

2.3 **Computed Tomography Reconstruction from the X-ray Image**

Several classic CT reconstruction methods, such as iterative reconstruction methods and filtered back projection [4] necessitate hundreds of X-rays data throughout a full rotational scan of the body. Lately, deep neural networks have also been applied to enhance the performance in this research area. For example, Würfl et al. [27] use the X-ray sonogram as input, which captures a volume from several different angles. They mapped Computed Tomography reconstruction to the back-propagation of a Deep Convolutional Neural Network (DCNN). Next, Hammernik et al. [28] proposed a deep neural network for a limited angle CT reconstruction. More close to our work is X2CT-GAN [2], which uses a generative adversarial network (GAN) to reconstruct CT volume from two corresponding biplanar X-ray images. Moreover, Bayat et al. [29] present a novel neural network called TransVert, which takes orthogonal 2D radiographs as input and concludes the spine's 3D posture. TransVert is comprised of sagittal and coronal 2D encoders with self-attention

modules and a 3D decoder. Single-image Tomography [30] utilizes an encoder-decoder based model to reconstruct 3D skull volumes scanned from 175 mammalian species from 2D carinal X-Ray images; however, the outcome is subject to too much ambiguity. Oral-3D [31] used a single panoramic X-ray image with some prior knowledge of oral structure to reconstruct Cone Beam Computed Tomography (CBCT). Kasten et al. [32] utilize an end-to-end deep CNN to reconstruct Computer Tomography (CT) from bi-planar X-ray images. In [59], Jiang et al. presented a deep convolutional neural network (DCNN) based volumetric image reconstruction method, allowing an end-to-end volume conjecture from 2D radiographs. Furthermore, they proposed a bi-directional cross-modal transition module to reconstruct 3D cone-beam CT (CBCT) images from 2D X-ray images.

Table 2. Several popular models to reconstruct CT volumes from different views of X-ray images input

Research Work	Method	Single View	Biplanar View	Multiple View
Herman et al. [4]	Iterative reconstruction methods and filtered back projection	×	×	✓
X2CT-GAN [2]	Generative Adversarial Network (GAN)	✓	✓	×
TransVert [29]	Encoder-decoder based method	×	✓	×
Single-image Tomography [30]	Deep convolutional neural network (DCNN)	✓	×	×
Oral-3D [31]	Encoder-decoder based method	✓	×	×
Kasten et al. [32]	End-to-end Deep convolutional neural network (DCNN)	×	✓	×
Jiang et al. [59]	Deep convolutional neural network (DCNN)	✓	×	×

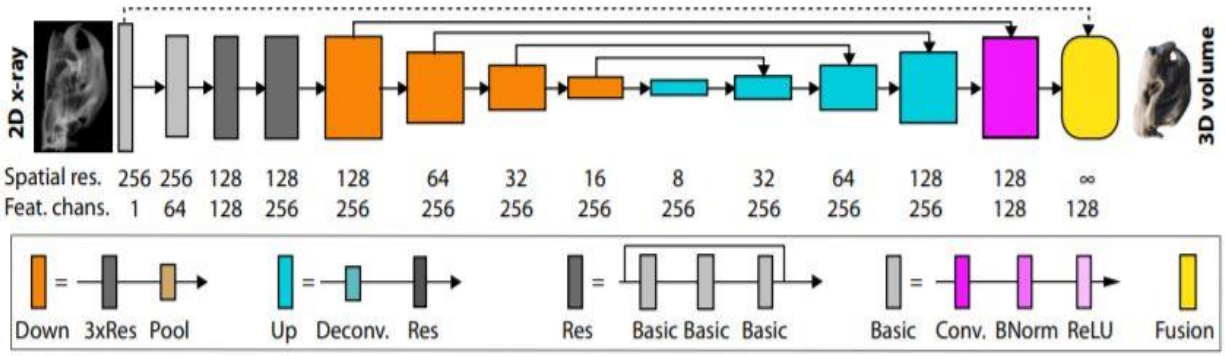

Figure 2-3. Design of the single image tomography model reprinted from [30].

Chapter 3 **Deep Learning Background**

3.1 **Convolutional Neural Networks (CNN)**

The convolutional neural network (CNN) is regarded as the most popular deep learning model [38]. Nowadays, CNN is widely used in image recognition, classification, object detection, image segmentation, etc. Convolutional neural network expands traditional neural networks by integrating both locally connected convolutional layers and fully connected hidden layers. Conventional neural networks impose high computational complexities by fully connected each neuron in the hidden layer to all neurons in the following layer. Moreover, this traditional neural network design process does not benefit the local correlations of pixels in practical-world images.

On the contrary, CNN construction is inspired by the biology of human vision, where it takes the information of the spatial correlation of objects present in practical-world images. Moreover, by developing an idea of the receptive field, CNN regulates each node's connectivity to a local region [39]. CNN is utilizing four basic concepts such as pooling, local connections, and efficiently employing many layers to tackle parameters and enhance the approximation accuracy.

3.1.1 **Convolutional Layer**

The convolutional layer is to extract filters from an input image in order to achieve useful patterns and features. In the deeper layer of the network, these filters learn more high-level features. At the initial stage of the training process, weights initialized randomly, but during the advanced stage of training, weights are adapted through back-propagation to learn features. These learned features are more indicative of accurate predictions.

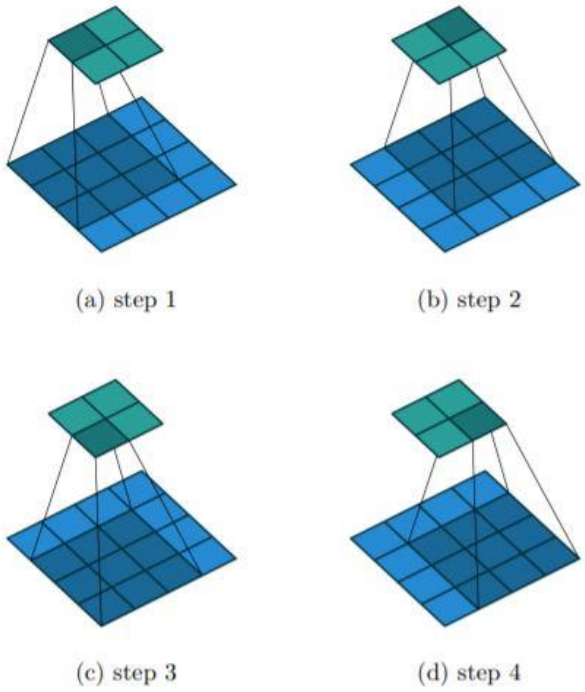

Figure 3-1. Convolution operation reprinted from [40].

Convolution conserves the correlation between pixels by learning image features utilizing small squares or windows of an input image. Through this window, we can inspect features within an image. This window also can be referred to as a filter or a kernel. Since it provides an outcome that emphasizes the area of the image that displayed the valuable features it was looking for. It is mathematical functioning that takes an image matrix and a kernel as inputs. The output of this convolutional operation is known as a feature map. If $an\ image\ matrix = h \times w \times d$ and $a\ filter = f_h \times f_w \times d$, then the convolution operation F can be formulated as:

$$F = (h - f_h + 1) \times (w - f_w + 1) \times 1 \qquad \text{Eq. 1}$$

Where h, w, and d can be denoted as image height, width, and dimension, respectively. Besides, f_h, and f_w can be mentioned as filter height and filter width.

In Figure 3-9, the input matrix (light blue) convolved by a kernel (dark blue) and originate a feature map (green). Here, zero-padding [41] has not been applied on the input matrix, and the $stride = 1$ on the kernel.

3.1.2 **Stride Convolutions**

Stride convolutions are a method for down-sampling the spatial resolution of our feature map. The stride length of a convolution determines how many steps should take during the sliding window (kernel) across an image. This sliding usually happens across the image from left to right. A standard convolution operation has a $stride = 1$, moving the kernel one unit in order to scan across the image. In contrast, we define $stride = 2$ during the downsampling process of our 2D encoder signifying that we move two units for each sliding (Figure 3-10.).

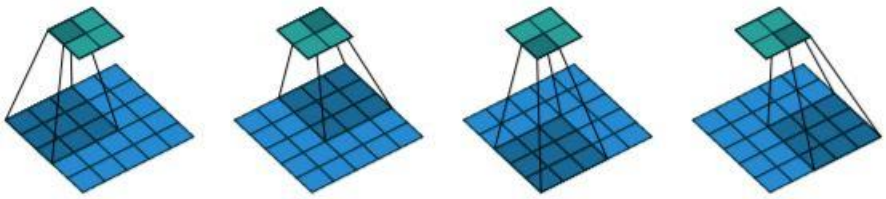

Figure 3-2. Convolution operation with kernel=3 and stride=2 reprinted from [40].

3.1.3 **Padding**

Padding is a term applied to the convolutional neural network (CNN) as it mentions the number of pixels attach to an image when it is being operated by the kernel or window of a convolutional neural network. If we take $padding = 0$ then the pixel in the corner of an image will only get notices one time. On the other hand, the pixel in the middle portion will get discovered more than one by the window. So, there will be some downsides if we do not use any padding:

1. shirking the output of our 3D decoder.
2. We are losing valuable information on the corners of our 3D volumes. As we tackle the issue of the medical domain, every detail is vital for our reconstruction task.

So we add some additional layers in the border of our 3D CT data when we run the convolution operation. Thus, now we have more frames that discover edge pixels more than once, and we have more probability of achieving an accurate result after this operation.

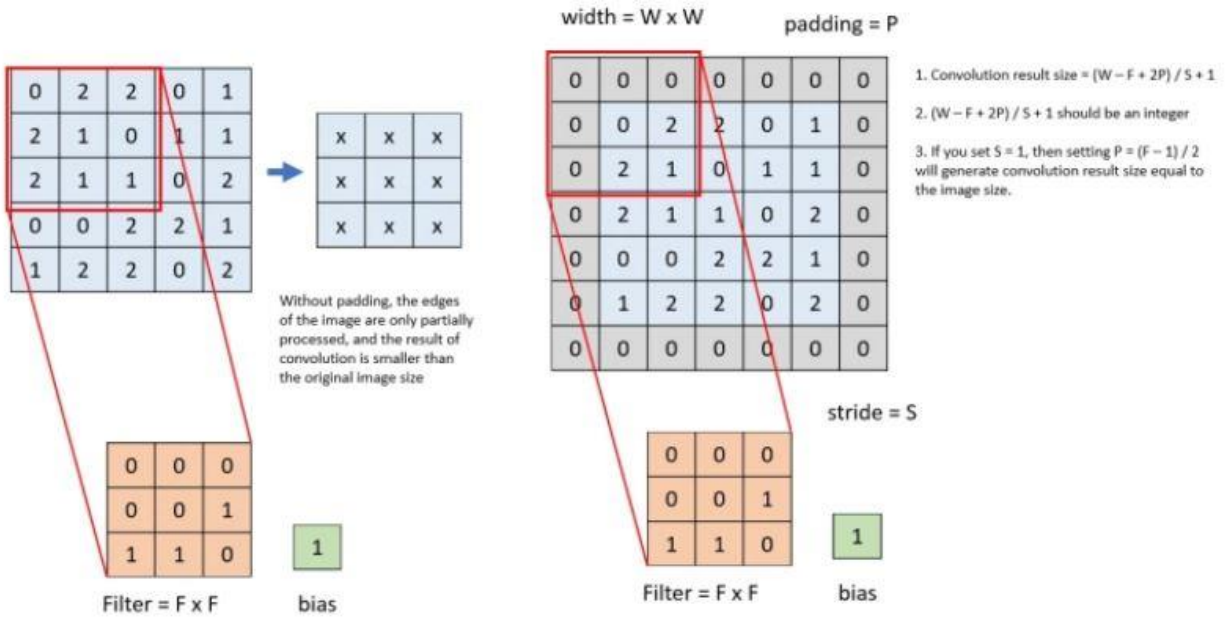

Figure 3-3. Convolution operation with padding on a feature map reprinted from [42].

3.2 **Activation Function**

The activation functions are simply a mathematical formula that regulates the outcome of a neural network. The activation function is employed to each neuron in the neural network and decided whether the neuron should be fired (activated). The activation function assists each neuron in normalizing a range between -1 and 1 or between 0 and 1. Another vital aspect of activation functions is that it processed across thousands or even millions of neurons to produce a computationally efficient system for each data sample.

The activation function is implemented on a neuron to locate some kind of non-linearity in the model. Without the activation function, the neural network will become a simple linear regression model that will not have the potentiality to solve complex problems. Therefore, to solve the complex issues, we need the activation functions in our network. Additionally, the non-linear activation function efficiently makes the model to generalize with a variety of data. So, we used two hugely popular activation functions, such as Rectified Linear Unit (ReLU) and LeakyReLU in our proposed CCX-rayNet.

3.2.1 **Rectified Linear Unit (ReLU) and LeakyReLU**

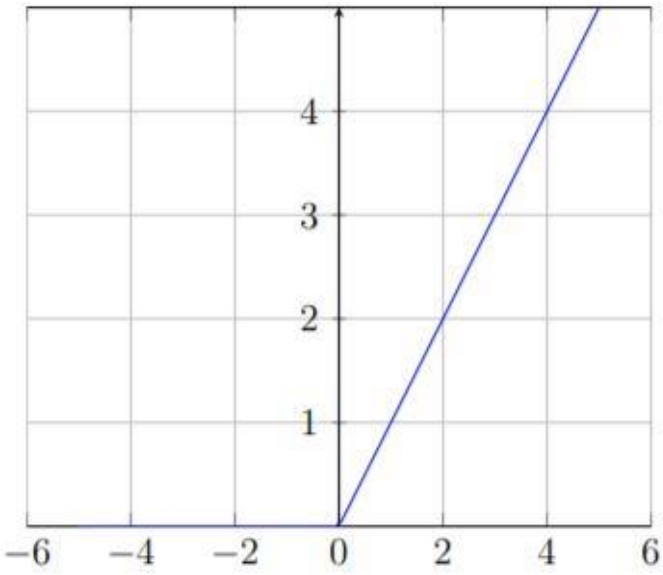

Figure 3-4. ReLU activation function reprinted from [45].

Rectified Linear Unit [43] or ReLU is the most popular activation function in the deep learning community. ReLU takes a real value as input, squashes the value between $0\ to + \infty$. ReLU avoids the vanishing gradient problem, allows the architecture to converge very swiftly, and perform computationally cheaper operations.

The ReLU activation function follows the following formula:

$$f(z) = \begin{cases} 0\ for\ z \leq 0 \\ z\ for\ z > 0 \end{cases} \qquad \text{Eq. 2}$$

Another way to write this equation as:

$$f(z) = \max(0, z) \qquad \text{Eq. 3}$$

However, ReLU suffers the dying ReLU problem, which means when inputs are negative or approach zero, this function's gradient becomes zero. At that time, the network can not execute

back-propagation, and the system loses the ability to train or fit from the dataset appropriately. Hence, it may not be acceptable for many datasets and models.

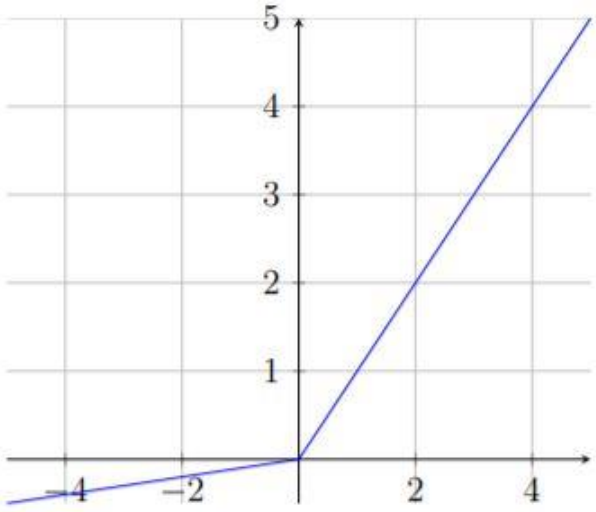

Figure 3-5. LeakyReLU activation function reprinted from [45].

So, Leaky Rectified Linear Unit or, LeakyReLU [44] might be an excellent solution to this complication. Instead of disregard all the negative activation, LeakyReLU registers a little positive slope (0.01 or so) in the negative region. Thus, the equation of LeakyReLU can be defined as:

$$f(z) = \begin{cases} 0.01z \; for \; z < 0 \\ z \; for \; z \geq 0 \end{cases} \qquad \text{Eq. 4}$$

Another way to write this formula as:

$$f(z) = \max(0.01z, z) \qquad \text{Eq. 5}$$

3.3 **DropOut**

Dropout is a regularization process for preventing overfitting in deep neural networks by reducing critical co-adaptations on training data. The dropout signifies randomly omitting or "dropping out"

neurons during the training of the network. By omitting, we do not consider these neurons during a particular forward or backward pass.

More technically, during the training phase, every node is either dropped out from the network with probability $1 - p$ or remains in the training process with probability p, so that a concise model can left.

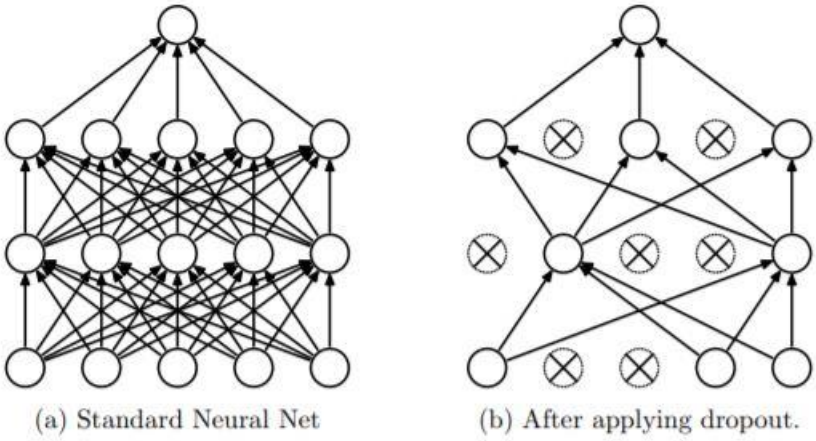

Figure 3-6. A dropout process of a standard neural network is reprinted from [46].

3.4 **Instance Normalization**

Normalization is a crucial factor to train a deep learning model effectively. Firstly, normalization makes the faster optimization so that the weights do not vary a vast range and restrict all weights inside a suitable range. Secondly, normalization minimizes the internal covariate changes so that it provides a faster training process. Next, it helps networks in the regularization process and also assists in handling the vanishing gradient problem.

There are several types of normalization, such as batch normalization, layer normalization, instance normalization, group normalization, etc. The success of a network vastly lies in selecting a suitable normalization to run the training process.

Recently, instance normalization [47] is extensively used in the image generator architecture (in encoder and decoder) because this normalization process permits to detach instance-specific knowledge from the content image, which produces a more straightforward generation process. Besides, this enhances the quality of the image. Next, instance normalization enumerates

mean/standard deviation and normalize each training example across each channel. Ulyanov et al. [47] replaced batch normalization with the instance normalization in their generator network, which helps them to avert intence-specific mean and covariance shift provides a simple learning process for the system. Moreover, it improves the performance of their deep neural architecture for image generation when they apply instance normalization in the place of batch normalization.

3.5 Generative Adversarial Networks (GAN)

Generative adversarial networks (GANs) are an exciting recent innovation of machine learning. Goodfellow et al. [3] designed this neural network back in 2014. Interestingly, Generative adversarial networks (GANs) are among one of the many types of generative methods that produce new sets of data from the training sets that spectacularly look like the original data.

A GAN contains two convolutional neural networks (CNN): generator and discriminator. The generator is trained to produce realistic-looking data based on the training data provided. At the same time, the discriminator tries to classify whether the newly generated data is a fake example of training data or not. The GAN is a kind of a feedback loop where the generator is assisting in training the discriminator, and similarly, the discriminator is assisting in training the generator.

Generative adversarial networks (GANs) resolve the issue of the inefficiency of utilizing per-pixel loss, depending on the adversarial neural network for feature detection, producing more effective loss values. The loss of discriminator is utilized to update the gradients of the generator to provide more realistic output images. Both generator and discriminator are linked so that the advancement in architecture helps training in the other until the discriminator can no longer precisely differentiate between generated and real images. A typical GAN architecture employs training data to achieve the intuition of data distribution, afterward try to use random noise input to create a random but realistic image that takes place within the feature space of the training data. In Figure 3-7, we displayed a typical generative adversarial network (GAN).

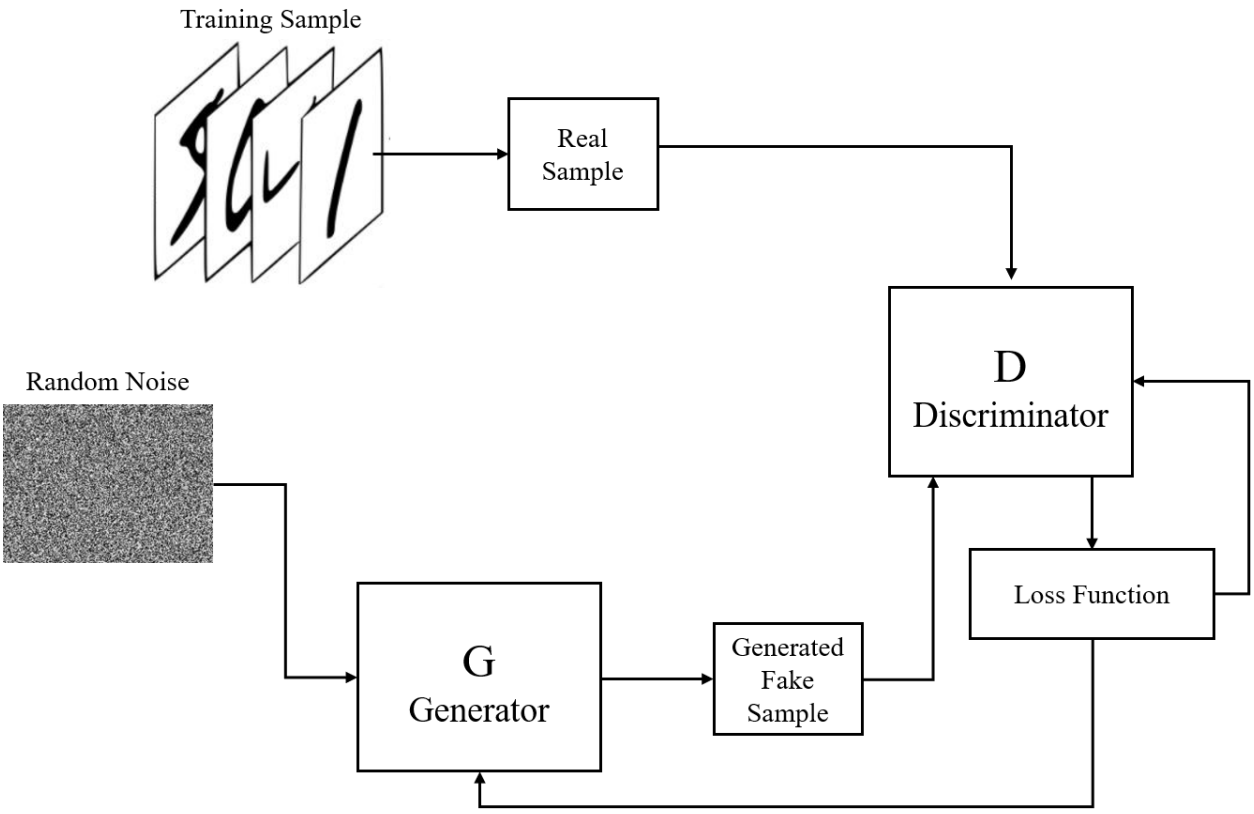

Figure 3-7. A typical Generative Adversarial Network (GAN).

3.6 **Encoder-Decoder**

The encoder is a network that takes an input, and the output is a feature vector/map/tensor. The feature map comprised the information or feature from the input. On the other hand, the decoder is a network that basically has a similar network arrangement as an encoder with opposite orientation. Furthermore, it takes the feature map from the encoder network and provides the best suitable match to the real input or intended output. In general, the encoder takes the input sequence and converts it to an internal representation, also known as 'context vector'. Then, the decoder decodes this information and produces the output sequence. The input and output sequence length can be different because there is no straightforward one to one relation between the input sequence and the output sequence.

Recently we observe several encoder-decoder networks for image reconstruction. Most of the networks stand for 2D to 2D image reconstruction and 2D to 3D image reconstruction. In this research work, our input is a 2D X-ray image, and the output is 3D CT volume. Therefore, we

design the encoder-decoder based network that contains a 2D encoder network and a 3D decoder network.

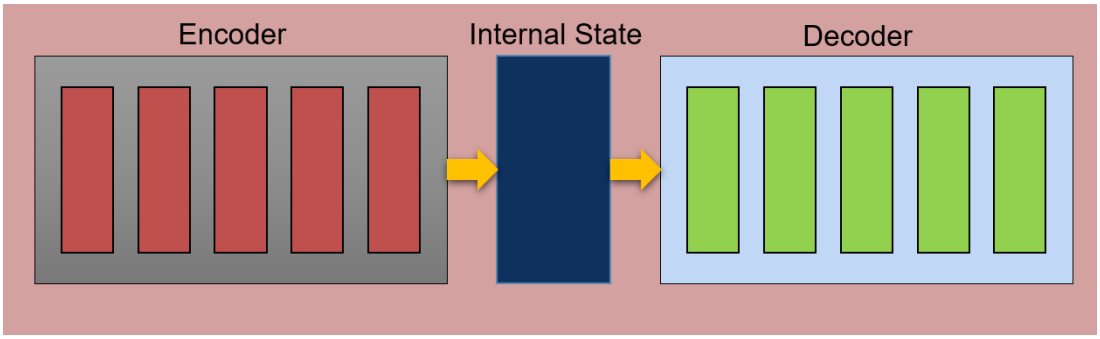

Figure 3-8. The basic encoder-decoder network.

3.7 Loss Function

The loss function is a method that evaluates how efficiently a prediction algorithm works to predict the desirable outcome. If the predictions diverge too much from real results, then the loss function would come up with a huge number because of inaccuracy. With the help of some optimization algorithms, loss functions discover to minimize the fault in the prediction.

Neural networks use the stochastic gradient descent optimization method in the training part. A specific part of the optimization method repeatedly estimates the error for the current state of the algorithm. During the configuration of the model, we necessitate a suitable loss function that can calculate the errors so that we can update the weights to minimize the loss in the next iteration. In machine learning, neural network algorithms learn a mapping from inputs to outputs from data, and the selection of loss function should match the framing of the relevant predictive modeling issue. The loss function can be divided into two main categories: Classification loss and Regression loss. It depends on the type of learning task we are performing.

- Classification: For the classification, we are trying to predict the outcome from a set of finite categorical values. Two well-known classification loss function:
    1. Hinge loss or Multiclass SVM Loss
    2. Cross-Entropy loss or Negative log-likelihood

- Regression: Regression based on predicting continuous values. Some well-known classification loss function:
    1. Mean Square Error or Quadratic loss or L2 loss

2. Mean Absolute Error or L1 loss
3. Mean Bias Error

3.7.1 Dice Coefficient Loss for U-Net

Dice coefficient loss is based on the dice coefficient, which is very much popular in the segmentation task. The dice coefficient computes the overlap between the two samples. The dice coefficient's value ranges from 0 to 1, where a dice coefficient of 1 indicates a complete overlap. This loss function initially establishes for the binary data and works better for the class imbalanced dataset, which is habitual in the medical domain. In a medical image segmentation task like ours, it is usual that the anatomy of interest inhabits only a tiny region of the image. Thus, the learning process often gets trapped into local minima, which makes network predictions biased towards the background. Besides, the vital foreground region is often only partially detected or get missing. To tackle this issue, Milletari et al. proposed the dice coefficient loss for their popular network V-Net [52]. However, we employ the dice coefficient loss in our U-net as this loss function works better for the medical domain.

$$Dice = \frac{2|X \cap Y|}{|X| + |Y|} \qquad \text{Eq. 6}$$

Here, $|X \cap Y|$ denotes the common elements between X and Y. In addition, $|X|$ and $|Y|$ respectively indicates the element inside the set X and Y. For the predicted segmentation masks, we can estimate $|X \cap Y|$ as the element-wise multiplication between the target and predicted mask, and then the summation of the resulting matrix. So, finally, for the loss function, we can formulate the dice coefficient loss as $1 - Dice$ because we need the predicted probabilities rather than thresholding and turning them into a binary mask.

3.8 Attention Mechanism

The attention mechanism achieves massive popularity in the NLP communities and in machine translation through Transformer Networks [60]. Lately, in computer vision, several research works

based on attention accomplished to complementing existing architectures or established some new ones.

The attention mechanism motivated us to pay visual attention to the different parts of the image in computer vision. Unlike natural photographs, medical images such as X-ray images are quite similar in their appearances. Medical images obtained the same acquisition parameters from different standardized positions. For medical personnel, the experience in examining the medical images comes mostly from knowing where precisely to focus on to discover a specific pathology. Therefore, it is quite understandable why attention plays a vital role in medical imaging.

By using the self-attention mechanism, Guan et al. [61] try to enhance the performance of an automated model that performs the classification of thorax disease from chest X-ray images. Previous classification models only designed to detect and execute classification on chest X-ray images (CXR), just focusing on the global image. The issue with utilizing the whole X-ray images for the classification task is that in medical image analysis, the lesion region may be so small compared to the whole image. Besides, the lesion may exist somewhere along the borders of the image. These problems produce a great amount of noise for the classifier and provide an inaccurate detection model. Additionally, chest X-ray images (CXR) are often suffering misalignment, which causes irregular borders in the images, and it negatively impacts the classification accuracy. So, tackling this issue, Guan et al. [61] proposed attention guided deep convolutional neural network where researchers first achieve the knowledge of global CNN from global images. Next, attention guided heat map generated from the global branch, and they surmise a mask to crop a discriminative area from the global image. Then, local area utilizing to train the local CNN branch of the proposed method. Lastly, they concatenated the last pooling layers of both the local CNN branch and the global CNN branch for the fine-tuning of the fusion CNN branch.

Chapter 4 **Methodology**

This section presents our proposed architecture design called CCX-rayNet with both single-mode and biplanar mode to accomplish 3D CT reconstruction tasks from 2D chest X-ray images (CXR). For both ways, we adopt the segmentation probability map to provide semantic information to the main network. Firstly, in the single view mode, we only consider the posterior-anterior (PA) view with the semantic segmentation maps to attain our work. Secondly, we consider both posterior-anterior (PA) and lateral view with the semantic segmentation maps to conduct our task for the biplanar view mode. We also provide a detailed explanation of the objective functions we use to train our proposed models.

In this work, we follow two approaches to perform our tasks:

1) The first one is the only encoder and decoder based generator network called CCX-rayNet. Besides, CCX-rayNet does not contain any discriminator network like a generative adversarial network (GAN) [3].
2) On the other hand, we proposed a GAN based approach, namely CCX-rayGAN, to raise the visual quality in our 3D volumes for the medical community. CCX-rayGAN is following the structure of other 3D GAN based network. This architecture incorporates a 3D generator and a 3D discriminator to generate our desired result.

These two different training procedures trained with the supervision explained in the “Objective Functionality of CCX-rayNet and CCX-rayGAN” section.

In these proposed models, we also include three new modules in the generator part:

1. Depth Feature Transform (DFT)
2. Depth Aware Connection (DAC)
3. Adaptive Feature Fusion (AFF)

Depth Feature Transform (DFT) and Depth Aware Connection (DAC) are used in both the single and the biplanar view approach. In contrast, Adaptive Feature Fusion (AFF) only considers during the biplanar view training process. Moreover, we acquire the segmentation probability map from

two U-Net [33] to equip our network with prior semantic information specifically for the lungs and heart. At first, we will give the intuition about our biplanar view method and then discuss the single view one.

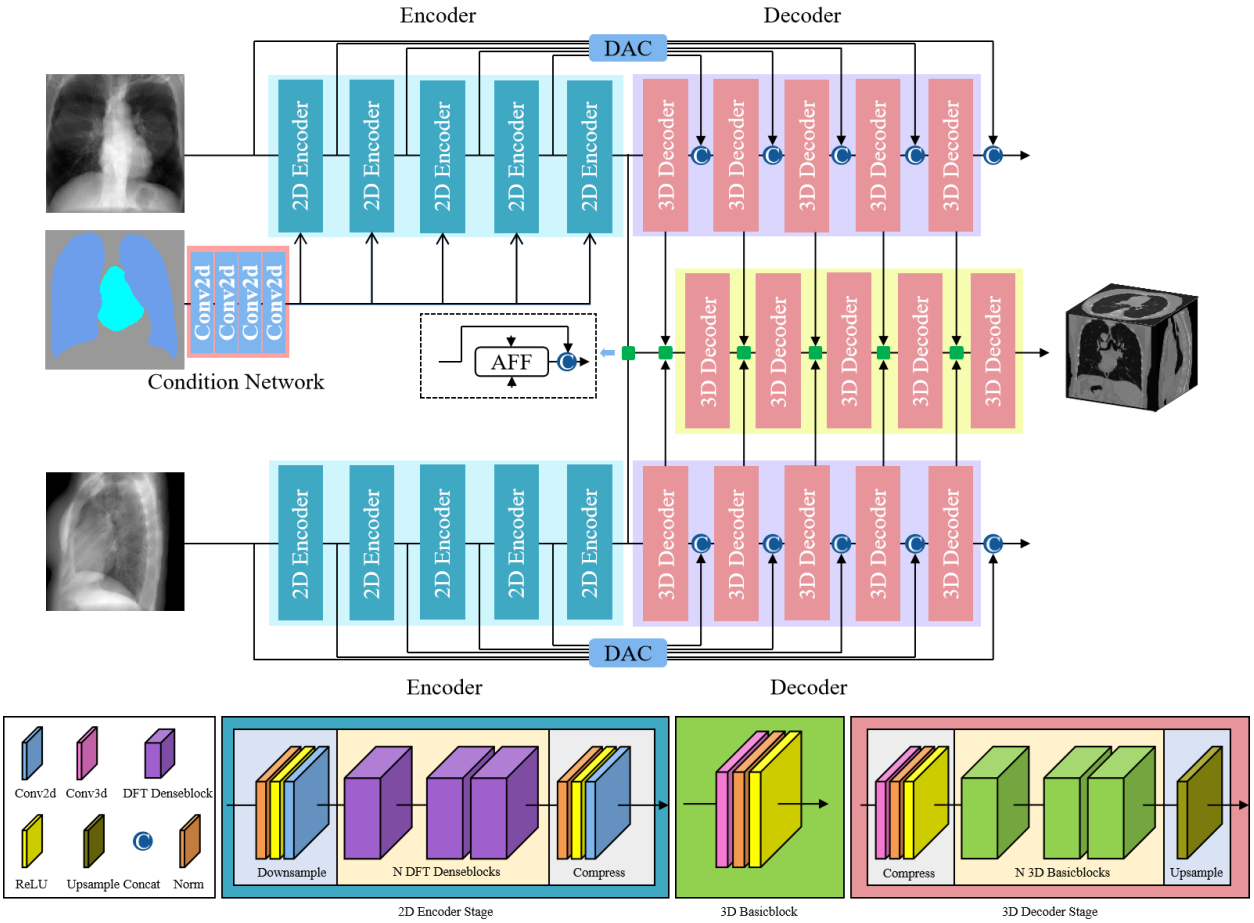

Figure 4-1. The generator part of our proposed biplanar CCX-rayNet with two encoder and decoder networks. Here, the inputs are posterior-anterior (PA) views of X-ray images with segmentation maps and lateral views of X-ray images. The model incorporates proposed DFT, DAC, and AFF modules.

4.1 Loss Functionality of CCX-rayNet and CCX-rayGAN

We train our model in two different manners: one with the reconstruction loss only where voxel-wise loss and projection pixel-wise L1 loss are combined, denoted as CCX-rayNet. In order to preserve the structural consistency between input and reconstructed results, we employ mean square error (MSE) loss as the reconstruction loss of the generator part. Furthermore, for the sharper image boundaries, we apply L1-loss as the projection loss.

To preserve anatomies, we train a GAN version of the model, called CCX-rayGAN. In the GAN based training approach, we set up the back-projection module of the generator and the discriminator by following a similar schema as a least-square loss like LS-GAN [34]. This training approach helps us to achieve perceptually feasible reconstruction and obtain realistic details in the resulting images.

4.1.1 **Least-square Loss (LS-GAN Loss)**

The main intention of GAN [3] to achieve the knowledge of latent data distribution by following an adversarial way to run the training process of the generator G and the discriminator D. The learning process of GAN is like a two-player game where a generator G and a discriminator D compete with each other. The generator expected to provides an output x_{fake} that looks indistinguishable from a real image x_{real} from a random vector z which sampled from a prior distribution $p(z)$. Moreover, the discriminator D plays the role of binary classifier here, which tries to distinguish between x_{real} and x_{fake}. Generate examples from a perfect generator can be indistinct from the real examples by the discriminator. In the min-max game, the generator adepts to estimated the real data distribution $P_{real}(x)$ through the optimization process. The following equation can mathematically summarize the min-max game:

$$\min_{G} \max_{D} V(G,D) = \mathbb{E}_{x\sim Pdata}[\log D(x)] + \mathbb{E}_{z\sim Pnoise}[\log(1 - D(G(z))) \qquad \text{Eq. 7}$$

Where z can be denoted as an example from a noise distribution.

From the semantic information of the X-ray image, we want to generate consistent 3D content. So, for the generator, we apply conditional GAN [35] to learn the cross-dimension mapping. However, after employing different mutants of conditional GAN, we found that to guarantee a stable training process and enhance the generation quality, LS-GAN is more appropriate to train the generator and discriminator. The conditional loss function of LS-GAN can be express as:

$$\mathcal{L}_{\text{LSGAN}}(D) = \frac{1}{2}[\mathbb{E}_{c\sim p(CT)}(D(c|x) - 1)^2 + \mathbb{E}_{x\sim p(X-ray)}(D(G(x)|x) - 0)^2] \quad \text{Eq. 8}$$

$$\mathcal{L}_{LSGAN}(G) = \frac{1}{2}[\mathbb{E}_{x\sim p(X-ray)}(D(G(x)|x) - 1)^2] \quad \text{Eq. 9}$$

Where x denotes the two orthogonal biplanar X-ray images, and corresponding CT volume is c. In equation 8, we use logarithmic loss, but in equation 9, we replace it with a least-square loss. In addition, due to this strategy, we achieve more realistic details in the CT volume and stabilize the adversarial learning process.

Moreover, the standard GAN network utilizes the sigmoid cross-entropy loss function, which causes a vanishing gradient problem during the training process. On the other hand, we use the least-squares loss function to overcome this issue.

4.1.2 **Reconstruction Loss**

In our CCX-rayNet we apply the mean square error (MSE) of the generation as the reconstruction loss. There are several benefits of using reconstruction loss:

- conserving the structural consistency with the input
- maintain the internal structure of the 3D volume
- reconstructed CT to be voxel-wise similar to the ground truth

Reconstruction loss of our generator can be formulated as:

$$\mathcal{L}_{\text{re}} = \mathbb{E}_{x,c}||c - G(x)||_2^2 \quad \text{Eq. 10}$$

Where x denotes the two orthogonal biplanar X-ray images, and the corresponding CT volume is c. The reconstruction loss is a voxel-wise loss that imposes the structural consistency in the 3D value.

4.1.3 **Projection Loss**

To enhance the training proficiency, we need a more exact shape prior that could be used as auxiliary regularizations. Inspired by [34], we take the 2D projections of our predicted volume to match the ones from corresponding 3D ground-truth in different views. We make this process straightforward by picking orthogonal projections instead of perspective projections. Furthermore, this auxiliary loss function emphasis only the consistency of general shape instead of the X-ray image's veracity. By following the medical imaging community, we select three orthogonal projection planes, such as axial, coronal, and sagittal views. Therefore, the proposed functionality for the projection loss can be determined as:

$$\mathcal{L}_{\text{pl}} = \frac{1}{3}[\mathbb{E}_{x,c}||P_a(c) - P_a\big(G(x)\big)||_1 + \mathbb{E}_{x,c}||P_c(c) - P_c\big(G(x)\big)||_1 + \mathbb{E}_{x,c}||P_s(c) - P_s\big(G(x)\big)||_1] \quad \text{Eq. 11}$$

In this formula, P_a, P_c, P_s represent respectively as view in the axial, coronal, and sagittal plane. Additionally, L1 distance is utilized to yield sharper image boundaries.

4.1.4 **The Combined Loss for CCX-rayNet**

To achieve the final loss function, we have to combine the reconstruction loss and projection loss. Thus, the combined loss of generator can be defined as below:

$$G^* = \arg\min_G [\mathcal{L}_{\text{re}} + \mathcal{L}_{\text{pl}}] \quad \text{Eq. 12}$$

Here, G represent generator, $\mathcal{L}_{\text{re}}$ stands for reconstruction loss, and $\mathcal{L}_{\text{pl}}$ denotes projection loss.

4.2 **3D Genarator of Biplanar CCX-rayNet**

We show the detailed architecture of biplanar CCX-rayNet in figure 4-1. The generator part of biplanar CCX-rayNet contains two encoder and two decoder networks with identical architectures,

one fusion network, and a condition network. The two encoder-decoder networks involve the 2D encoders and 3D decoders. Additionally, these two architecture work, respectively for posterior-anterior and lateral X-ray images. The fusion network contains only 3D decoders, and the condition network includes four convolutional 2D layers.

The encoder-decoder networks try to achieve the knowledge from the input 2D chest X-ray and the target 3D CT volumes in feature space. The encoder reduces a little resolution of features while enhances the number of feature channels. In contrast, the decoder raises feature resolutions. The 3D fusion network fused the biplanar information from these two networks and managed to produce the 3D CT scan. Finally, the condition network responsible for producing conditions from the 2D semantic map of frontal view X-ray images and provides this information of shape and position to the 2D encoders. The main task of our training process to disseminate knowledge of two distinct dimensionalities and modalities. So, we apply several modifications in our proposed architectures so that it can appropriate for the challenges of cross-dimensionalities and cross-modalities.

4.2.1 **Condition Network**

The condition network (shown in Figure 4-2) takes the semantic segmentation map as the input. The semantic map yielded from the posterior-anterior view of X-ray images and gave us valuable information about shapes, structures, and positions, particularly for the heart and lungs. So, we dispense this information to that encoder-decoder network, which responsible for the frontal view. The condition network has four convolutional 2D (Conv2D) layers with the LeakyRelu activation function. Furthermore, four Conv2D layers of condition network respectively have 128, 128, 128, 64 channels, or hidden units. The input dimension of the semantic segmentation maps is (128×128).

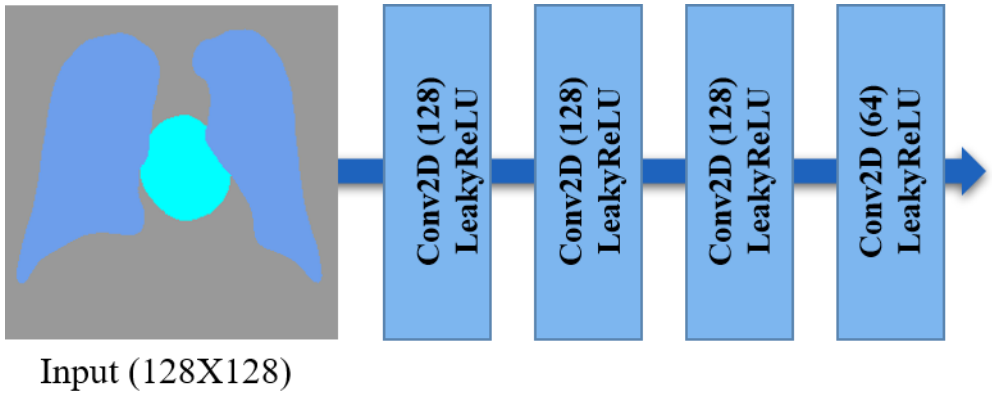

Figure 4-2. Structure of our condition network.

4.2.2 **2D Encoder**

The CCX-rayNet contains five 2D encoder modules in the encoder-decoder networks. Each encoder includes one Downsample module, a Densely connected DFT module, and a compression part. This densely connected encoder grasp a different level of vital information from input 2D X-rays and pass this knowledge to the 3D decoder along through Depth-Aware Connection (DAC).

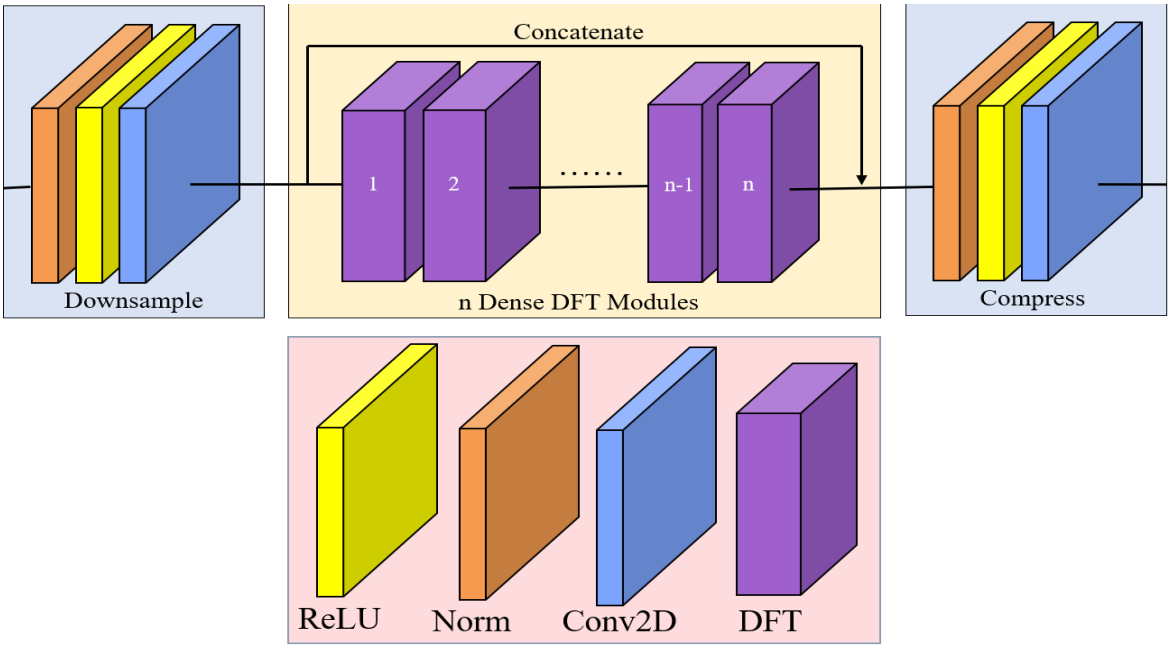

Figure 4-3. Design of a 2D encoder module.

4.2.2.1 **Downsample Module:**

The downsampling module comprises three layers: instance normalization, rectified linear unit (ReLU) activation functionality, and a 2D convolutional layer (Conv2D). This layer reduces the dimensionality of the features, which produce a computationally inexpensive encoder. Moreover, we can only focus on essential information that helps the network to minimizes errors. The Conv2D layers are implemented with kernel size 3×3 and $stride = 2$ where each Conv2D layer of five encoders hold 64, 128, 256, 512, and 512 channels, respectively.

4.2.2.2 **Deep Feature Transform (DFT):**

To better preserve the topology and shape information, we add a constraint to guide the network with semantic information. As shown in Figure 4-1, the segmentation map of the frontal view X-ray is fed into the condition network to generate the conditions. The DFT module will produce a modulation parameter pair (α, β) to transform the 2D feature. Specifically, those two parameters are estimated by a mapping function M whose input is the conditions,

$$(\alpha, \beta) = M(c) \qquad \text{Eq. 13}$$

Where c denotes the input conditions. After obtaining the parameter pairs, the original feature map will be scaled and shifted by the following equation:

$$F_{out} = DFT(F_{in}) = F_{in} \odot \alpha + \beta \qquad \text{Eq. 14}$$

Where the dimension of F_{in} is the same as α and β. This promises that the element-wise product $\odot$ and element-wise addition (+) operations are feasible, leading the DFT module to perform an element-wise transformation conditioned on the specific semantic prior.

Figure 4-4 shows the DFT module's details, where the mapping function M is modeled by four neural layers. In this work, M is optimized by end-to-end training. To improve the efficiency of the DFT, we set the conditions as the shared intermediate values for all DFT modules. As shown

in Figure 4-1, conditions are broadcasted to every 2D encoder stage so that the DFT module only possesses few parameters.

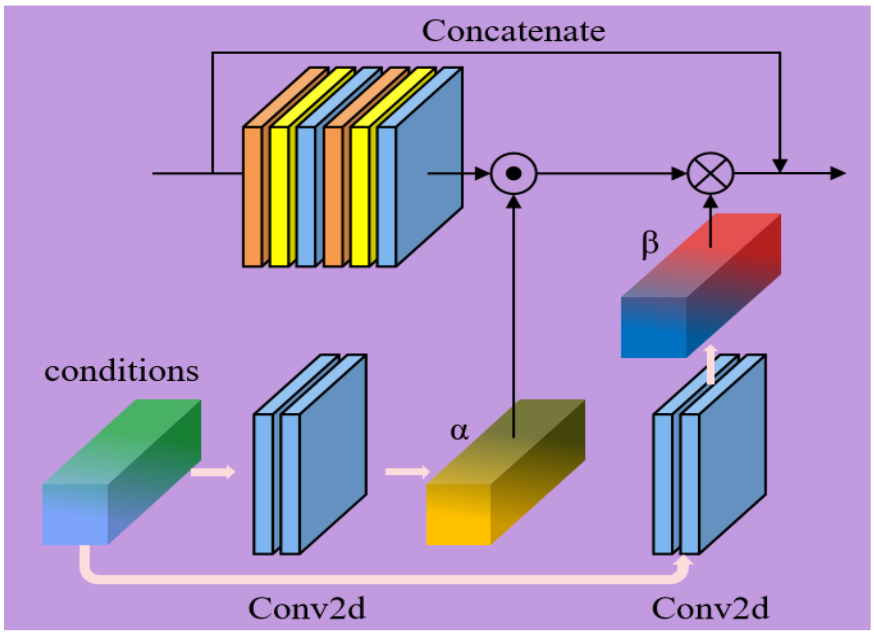

Figure 4-4. Structure of DFT DenseBlock.

We apply the skip connection [55] mechanism in the DFT module that means the final feature map F from the DFT module will be:

$$F = F_{in} + F_{out} \qquad \text{Eq. 15}$$

Here, F_{in} is the input feature information of the DFT layer and F_{out} achieved from Eq. 15. We implement a different number of dense DFT modules in the separate 2D encoder. The first and the last 2D encoder contains 6 DFT modules, whereas the second, third, and fourth 2D encoder include 12, 24, and 16 DFT modules, respectively.

4.2.2.3 **Compression Part**

Every compression part contains instance normalization, ReLU activation function, and a Conv2D layer with kernel size 1×1 and $stride = 1$. The first two encoders have 128, and 256 hidden

units and channels in the Conv2D layer in their compression part, respectively, whereas the third and fourth 2D encoder has 512 channels in the convolutional layer. Moreover, the convolutional layer of the last encoder comprises 352 channels in the compression part.

4.2.3 Depth-Aware Connection (DAC)

In this section, we provide an illustration of the inspiration of DAC and how the voxel-wise attention mechanism is applied. Recently, some encoder-decoder based networks are proposed to tackle the 3D reconstruction problem, but their pure connection design is losing the depth information in the perspective projection and not specially designed for this clinical scenario. Inspired by [2], we design a skip connection between 2D and 3D feature maps with the consideration of depth. According to the principle of X-ray image generation, DAC can complete the information lost in the compressed axis.

Considering the X-ray is the perspective projection of the CT volume, voxels are overlaying and compressed. Each slice of CT exhibits different information. Take the frontal view as an example, each slice in CT only contains subset information of the X-ray. However, expanding operation makes each slice the same feature and ignores the inter slices depth information. This inspires us to develop a mechanism to highlight certain areas from X-ray images for different slices of CT, which we call the depth-aware connection (DAC).

The DAC takes the following steps: 1) expand the 2D feature map to an intermediate 3D one by duplicating the 2D feature map along the third axis; 2) feed the 3D feature map into a basic 3D convolution block to enforce it with the shape of (B, C, D, H, W); 3) generate an attention matrix with the shape of (B, D, H, W) from 2D feature map; 4) expand the attention matrix C times and element-wise multiply it with 3D feature map. The above steps could be summarized as:

$$G = Conv2D(I) \quad \text{Eq. 16}$$

$$A = sigmoid(E(Conv2D(I))) \quad \text{Eq. 17}$$

$$\widehat{\mathrm{G}} = \mathrm{Conv3D}(\mathrm{G}) \quad \text{Eq. 18}$$

$$DAC(I) = A \odot \widehat{\mathrm{G}} \quad \text{Eq. 19}$$

Where I denotes the input 2D feature map, G, and $\widehat{\mathrm{G}}$ represent the intermediate 3D feature map and modified 3D feature map before adding the attention map. The attention matrix is generated from the I, and the expanding operation is referred to as E.

4.3 Connection Between 2D Encoder and 3D Decoder

Many existing popular encoder-decoder methods [12] [36] used convolution to link the encoder and decoder part where they work with the same dimensionalities. However, our approach is totally different in this work as our network tries to establish a connection between different dimensionalities. We utilize two bridging strategies to connect the 2D encoder and 3D decoder.

4.3.1 1st Connection

At first, to establish the first connection, we get inspiration from [37]. In the middle of our generator network, we apply the fully connected layer to link the 2D encoder and 3D decoder to set up the first connection mechanism. Precisely, through the fully connected (FC) layers, 1st connection attains the 2D to 3D conversion. Here, the result of the last encoder layer flattened and then broadened to a one-dimensional vector. Furthermore, we reshaped this one-dimensional vector to 3D. This connection layer holds instance normalization, a fully connected layer, and ReLU activation with a dropout rate of 0.50. Finally, we apply the 1st connection to bridge the last 2D encoder layer to the first 3D decoder layer.

Next, to prevent the loss of essential 2D spatial information during this conversion process, we proposed a new feature fusion method called Adaptive Feature Fusion (AFF).

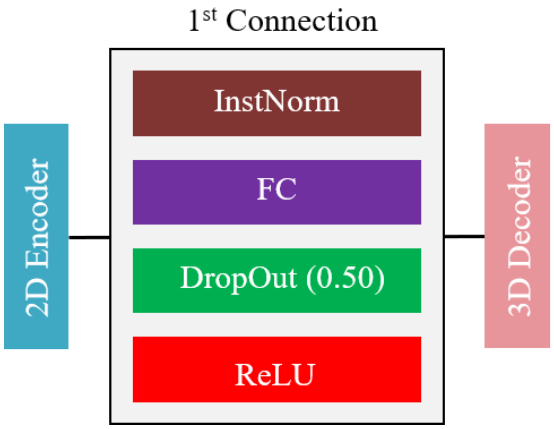

Figure 4-5. Structure of the 1st connection.

4.3.2 **Adaptive Feature Fusion (AFF)**

As mentioned previously, Ying et al. [2] propose a training set with synthesized orthogonal views of X-ray images, which is unrealistic in the real world. Towards the real clinic X-ray image acquisition process, biplanar X-ray images are offset with each other in a small margin. Yan et al. [56] induced spherical topology on the point-cloud to achieve the rotation-invariant convolutional neural network. However, it is designed exclusively for the classification task, and the computation burden can not be ignored in our CT reconstruction task. Therefore, we adopt a similarity matrix and attention mechanism to adaptively fuse the feature from orthogonal views efficiently, named AFF module.

To fuse two 3D feature maps, we compute a similarity matrix SIM to store the distance of two feature maps.

$$SIM = I_1 \odot I_2 \qquad \text{Eq. 20}$$

Where I_1 and I_2 are the orthogonal feature maps. SIM represents the distance between two corresponding feature voxels in neural space. In particular, we adopt the dot distance as the distance measurement since it is computational friendly. Noted that the similarity matrix only presents if the corresponding feature voxels are consistent, unlike the non-local module [57], which computes the distance between all voxel pairs. Our similarity matrix can be considered as a typical example of the non-local module.

Since we know alignment information, the network can adaptively enhance one of the feature points and compress the other one's contribution by generating two weighting matrices.

$$W_1 = Conv3D(SIM) \quad \text{Eq. 21}$$

$$W_2 = Conv3D(SIM) \quad \text{Eq. 22}$$

Where the $Conv3D$ operations do not share the weights. Finally, the weighted sum is operated to fuse the features by multiplying the 3D feature maps with their corresponding weighting matrices.

$$AFF(I_1, I_2) = W_1 \odot I_1 + W_2 \odot I_2 \quad \text{Eq. 23}$$

4.4 3D Decoder

Similarly, as the 2D encoder, we have a total of five 3D decoders in the decoder part of CCX-rayNet, where the first four 3D decoders have the same structure. Moreover, each of these four decoders contains three sections: compress part, N 3D basic blocks, and upsampling.

Firstly, the compress layers include one convolutional 3D layer ($Conv3D$) and one instance normalization layer with the ReLU activation function, $padding = 1$, and $kernel\ size = 3$. The number of channels is 16, 32, 64, and 128, respectively, for the four compression layers of the first four 3D decoders.

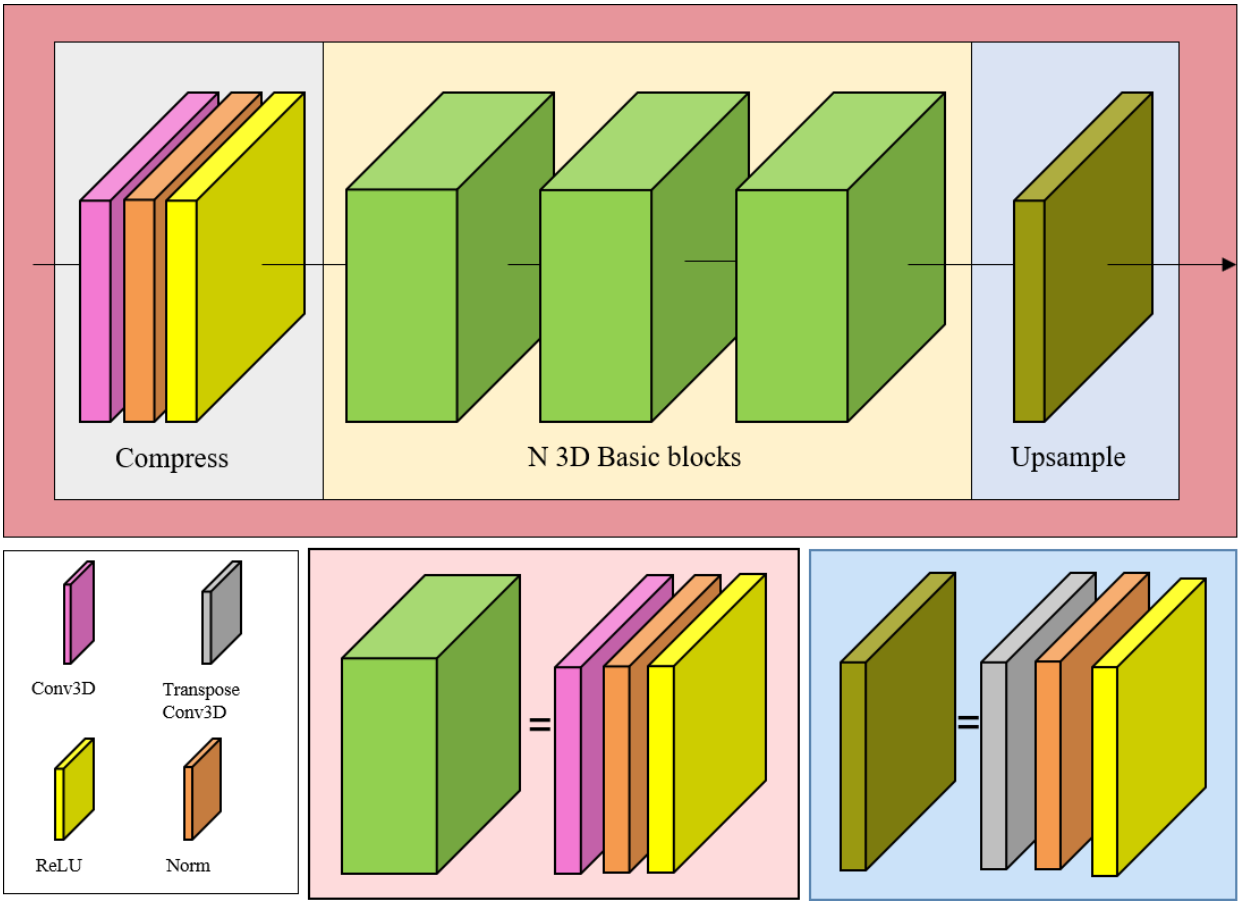

Figure 4-6. 3D decoder design of CCX-rayNet.

Secondly, we have three basic 3D blocks (N=3) in these four decoder layers. Furthermore, every basic block from a particular decoder has the same building structure. Individual 3D block contains a convolutional 3D layer ($Conv3D$) and an instance normalization layer with the ReLU activation function, $padding = 1$, and $kernel\ size = 3$. The hidden units or filters number in the $Conv3D$ of 3D basic blocks are 16, 32, 64, and 128, respectively, for the four decoders.

Finally, we use transpose convolution for the up-sampling block of the decoders. Transpose convolution or deconvolution layers apply to the decoder to enhance the resolution of the feature maps. We employ a 3D transpose convolutional layer with $stride = 2$, $kernel\ size = 3$, and $padding = 1$. Furthermore, the upsampling layer comprised an instance normalization layer with ReLU activation. The 3D transpose convolution involves 16, 16, 32, and 64 filters or channels in the four decoder layers.

The structure of the 5th decoder is a little bit different than the others. It has a compress layer and three basic 3D blocks like previously, though it does not include any upsampling layer. Instead of an upsampling layer, it comprises a 3D convolutional layer with $channel = 1$, $padding = 3$, $kernel\ size = 7$, and ReLU functionality. The compress layer and basic 3D blocks accommodate the same parameters as previous decoders, and both hold 16 channels.

4.5 Fusion Network for Biplanar View

We parallelly trained two networks for frontal view and lateral view X-ray images. Hence, it is easy to comprehend that the information generates by the two encoder-decoder systems is not the same. However, to achieve high visual quality in our output and trained our generative model, we need this complementary information from two parallel networks. So, here we use a 3D decoder network (fusion model) to fuse the extracted features information from each view, which we obtain during the parallel training process of the two encoder-decoder models. Finally, we acquire the 3D reconstructed result from this network. This fusion network is only applicable for our biplanar view modes because the main and the only work of this network is to fuse the valuable information which we gather from two parallel architectures. Thus, in our proposed single view architectures, we do not have any fusion network.

In Figure 4-7., we try to provide the intuition behind every 3D decoder layer of the fusion network. The fusion network will perceive any structural inconsistency between two 3D decoder networks and then back-propagated to two encoder-decoder models.

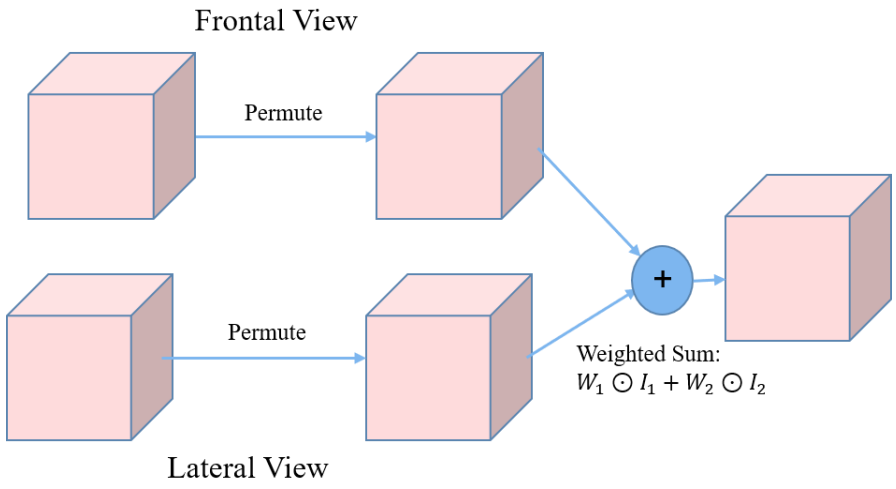

Figure 4-7. Fused outcome of two different views.

4.6 **3D Generator of Single View CCX-rayNet**

We display the detailed architecture of the single view CCX-rayNet in Figure 4-8. The generator part of the single view CCX-rayNet contains one encoder and one decoder network, and a condition network. Moreover, CCX-rayNet does not include any fusion network, so we do not employ any AFF modules for this model. The encoder-decoder system involves the 2D encoders and 3D decoders. Furthermore, this single view model only works with the posterior-anterior view of the X-ray image. We already discuss the structure and fundamental building blocks of the encoder-decoder network. However, the decoder part of single view mode only uses an extra 3D decoder layer in the decoder part. Finally, our GAN based single view model (CCX-rayGAN+S) utilizing the same 3D discriminator like biplanar mode.

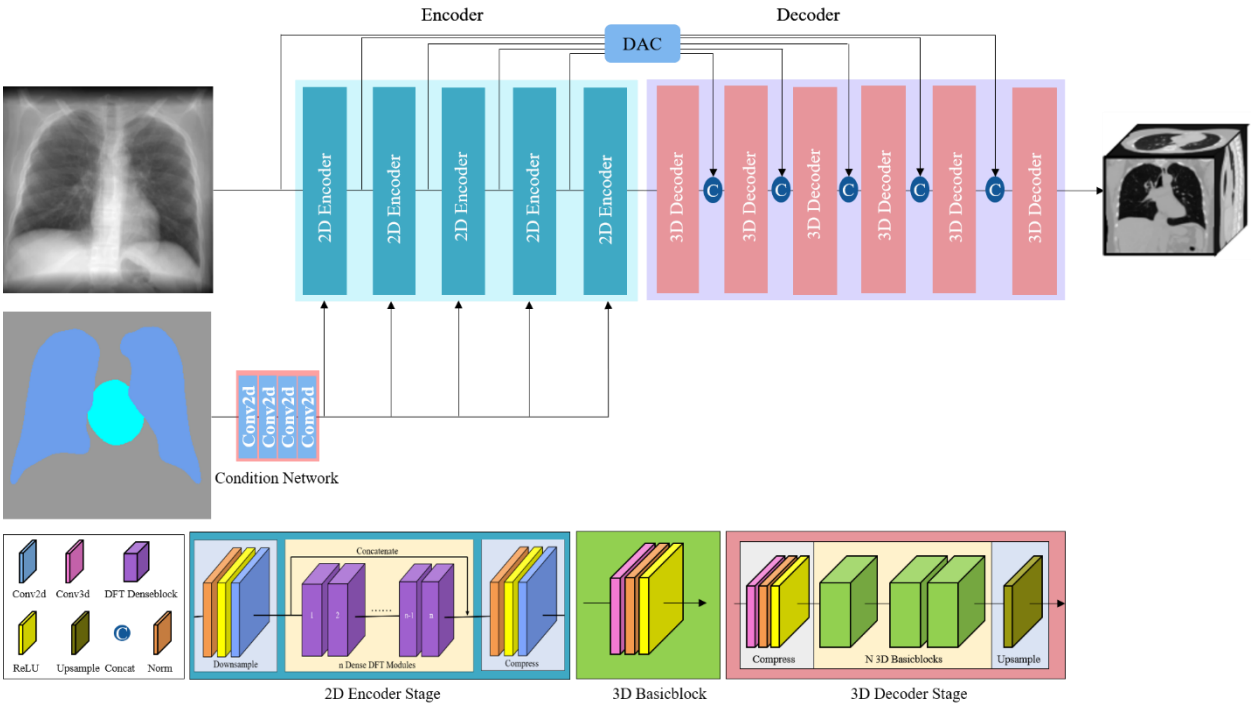

Figure 4-8. The structure of our proposed single view CCX-rayNet (CCX-rayNet+S).

4.7 **Discriminator**

A discriminator is simply a classifier in a generative adversarial network (GAN). The discriminator model tries to differentiate actual data from the data generated by the generator network. The actual or real data comes from the dataset. During the training phase, the generator tries to produce images like real data. After some time in the training process, the discriminator can not differentiate between the real and the fake images. The discriminator assists the generator in improving the image quality by upgrading the classification accuracy for real and fake images.

Lately, PatchGANs were utilized in many research work due to the adequate generalization properties of these architectures. We also employ 3D Patch Discriminator [12] in CCX-rayNet to improve the visual quality of 3D volume. Our 3D discriminator includes three modules:

1st module: $Conv3D - InstNorm - ReLU$ with $kernel\ size = 4$ and $stride = 2$,

2nd module: $Conv3D - InstNorm - ReLU$ with $kernel\ size = 4$ and $stride = 1$,

3rd module: one $Conv3D$ layer

Here, $Conv3D$, $InstNorm$, and $ReLU$ stands for convolutional 3D layer, instance normalization, and rectified linear unit, respectively. Our 3D discriminator network enhances the discriminative ability and can differentiate between real and fake 3D volumes.

Chapter 5 **Dataset**

To run the empirical analysis of this CT reconstruction method, we need a large dataset with the X-ray image and its corresponding CT scan. To produce a precise projection matrix X-ray machine requires to be calibrated correctly. As we need a CT and the corresponding X-ray image paired dataset, but we do not find any dataset which can fulfill our requirement. It is so costly to produce a real paired dataset. We used publically available LIDC-IDRI [48] datasets to create synthetic X-ray images to run this experiment. Specifically, this dataset contains 1018 chest CT scans.

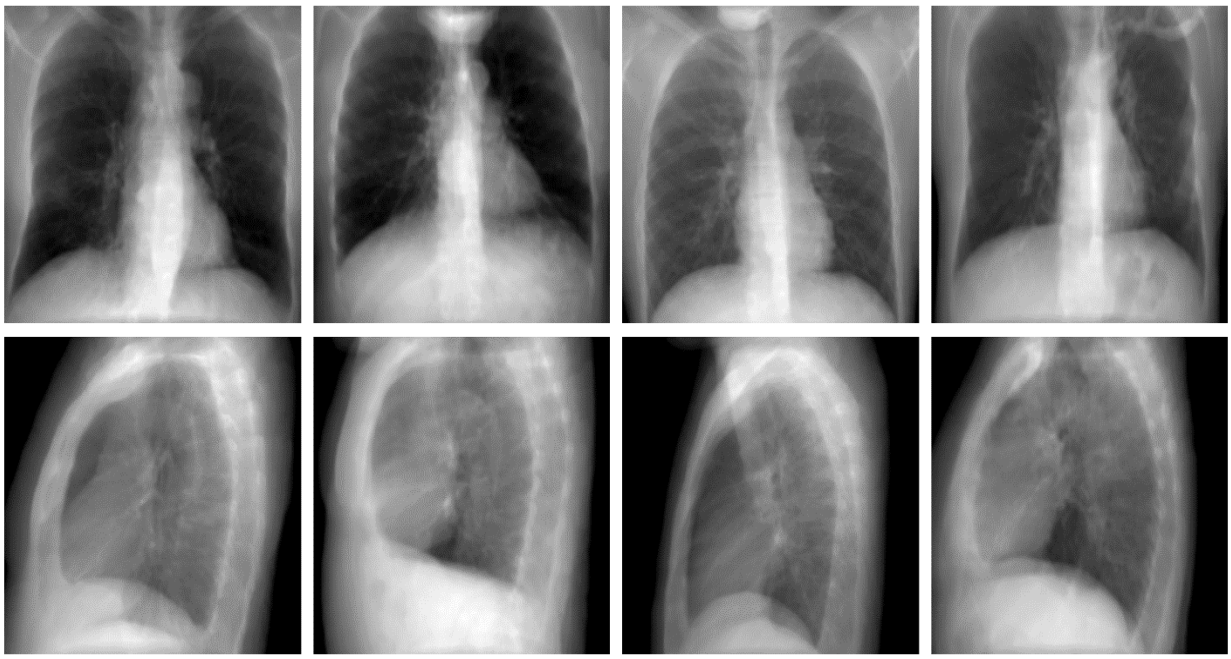

Figure 5-1. Some examples of our generated synthesized X-ray image data. First Row: posterior-anterior (PA) view, Second Row: Lateral view.

5.1 **Data Pre-Processing**

In this part, we provide a detailed explanation of an augmented dataset built on the LIDC-IDRI dataset. Additionally, we demonstrated a step by step synthesized dataset making process. Next, we explain the input and output images dimension of our networks, which includes X-ray images,

segmentation maps, and 3D CT volumes. Then, we execute a data standardization process for X-ray images and CT data. The main reason behind this standardization process is to use a standard scale for the X-ray images and CT voxels so that we can avoid a massive range of values in our data or losing information.

5.1.1 Procedure to Make Paired Dataset

We follow two steps to produce this dataset:

- We apply a digitally reconstructed radiograph (DRR) method to provide the X-ray images, but it missed some subtle information missing, such as blood vessels and lungs veins.
- Then we employ the CycleGAN method to produce the synthetic X-ray image data (close to real X-ray images).

5.1.1.1 Digitally Reconstructed Radiograph (DRR)

In order to provide the paired dataset, we follow the procedure of X2CT-GAN [2]. At first, we utilize an algorithm, namely, digitally reconstructed radiograph (DRR) calculation, to synthesize corresponding X-ray images from real CT volume. To be specific, to train and validate our proposed network, we use 1018 chest CT scans from the LIDC-IDRI dataset [48] with the corresponding synthetic X-ray images. The LIDC-IDRI dataset contains different numbers of CT slices for each CT scan. The resolution varies for various volumes, however the resolution is isotropic inside a slice. Specifically, these two aspects lead to a complex reconstruction task from X-rays, but to make it simple, we resample the CT scan to the $1\,X\,1\,X\,1\,mm^3$ resolution. Furthermore, we cropped $320\,\times\,320\,\times\,320\,mm^3$ the cubic area from each CT volume.

We found that digitally reconstructed radiograph (DRR) produce quite photo realistic X-ray images, but there is still some subtle information missing, such as blood vessels and lungs veins. Therefore, we need an image-to-image translation method that can help us transform an image from one domain to another. For example, here, we fancy to creating a more realistic synthetic X-ray image from DRR produced X-ray image data.

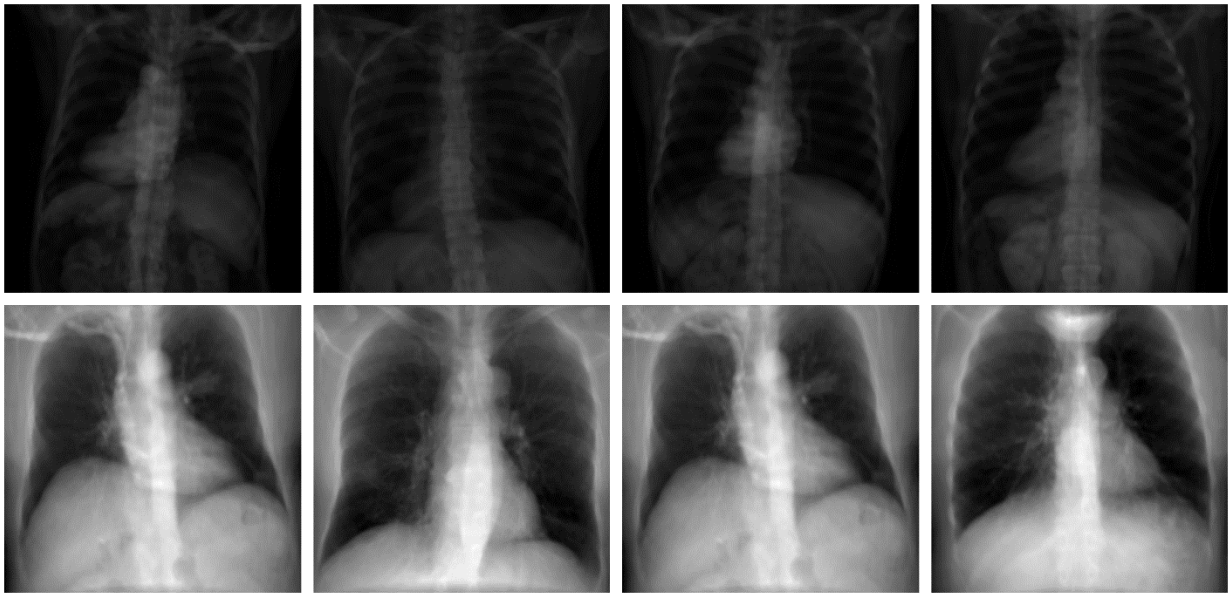

Figure 5-2. Few examples of posterior-anterior (PA) view of X-ray images. First row: posterior-anterior (PA) view of X-ray images produced by the digitally reconstructed radiograph (DRR). Second row: Style transfer methodology of CycleGAN used to make these X-ray images more realistic.

5.1.1.2 **Cycle-GAN Provided Synthesized X-ray Images**

The primary goal of the image-to-image translation method is to discover the mapping between an input and an output image utilizing a training set of aligned image pairs. However, it is not always feasible to obtain the paired images. CycleGAN [13] attempts to grasp the knowledge of mapping without requisite paired input-output images as it is a cycle consistent adversarial networks. CycleGAN comprised two Generative Adversarial Models (GANs) (total of two generators and two discriminators). We utilize the style transfer technique with CycleGAN to learn the mapping of real X-ray images and provide more realistic synthetic X-ray image data.

Later in our created dataset, we append predicted semantic segmentation maps for each synthetic X-ray images. In the dataset, each data contain a posterior-anterior (PA) view of the X-ray image, lateral view of the X-ray image, semantic segmentation map of heart and lungs, and corresponding CT value.

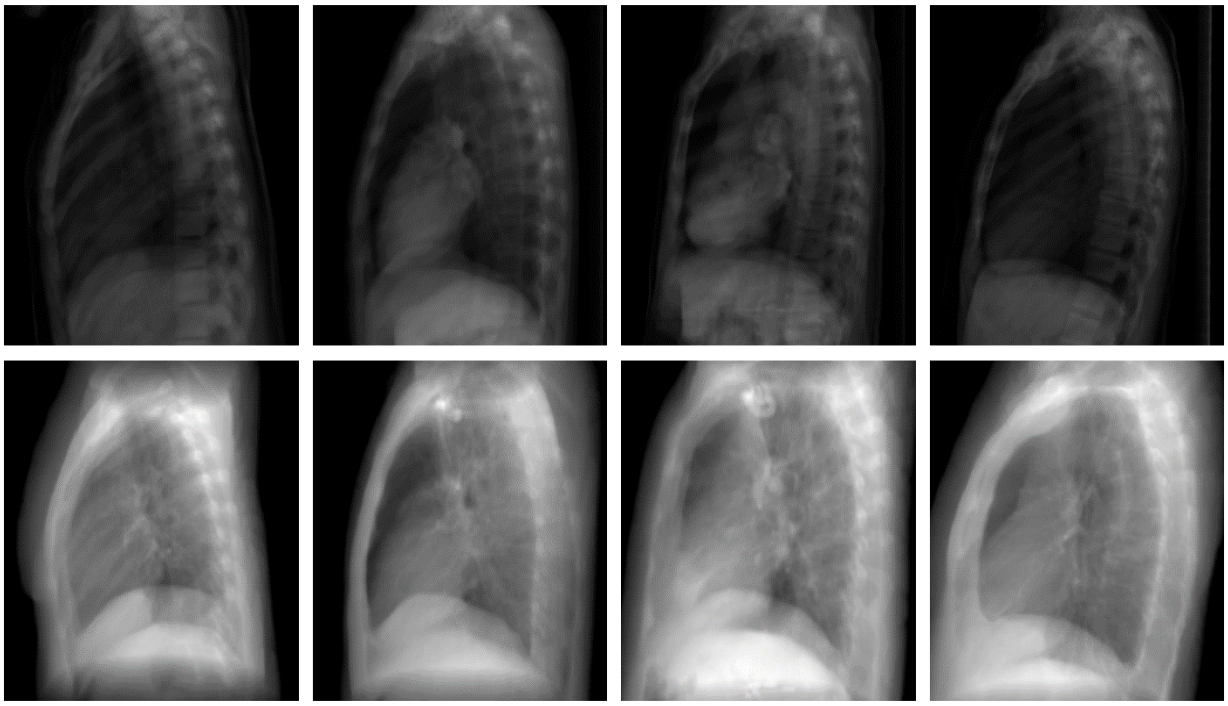

Figure 5-3. Few examples of the lateral view of X-ray images. First row: lateral view of X-ray images produced by the digitally reconstructed radiograph (DRR). Second row: Style transfer methodology of CycleGAN used to make these X-ray images more realistic.

5.1.2 **Segmentation Map**

The main goal of semantic image segmentation is to label each pixel of a particular image with the corresponding categories of what is being depicted. In semantic segmentation, we are predicting every pixel in the image, which is very often denoted as the dense prediction. In semantic segmentation tasks, we only think about the category of each pixel of an image rather than separating examples of the same class. The outcome of semantic image segmentation is a high-resolution image that can be referred to as a segmentation map. Generally, the dimension of a segmentation map is the identical size as the input image.

We adopt the segmentation probability map to provide semantic information to the main network. The primary reason is to provide some detailed information on some major organs such as the lungs and heart to the main network. Here, we give a brief explanation of the segmentation network we used. We train two binary U-Nets [33] for lungs and heart segmentation. Our main target is to

use a grayscale posterior-anterior (PA) chest radiographs ($height \times width \times 1$) and provide a segmentation map where each pixel holds a class label designated as an integer ($height \times width \times 1$).

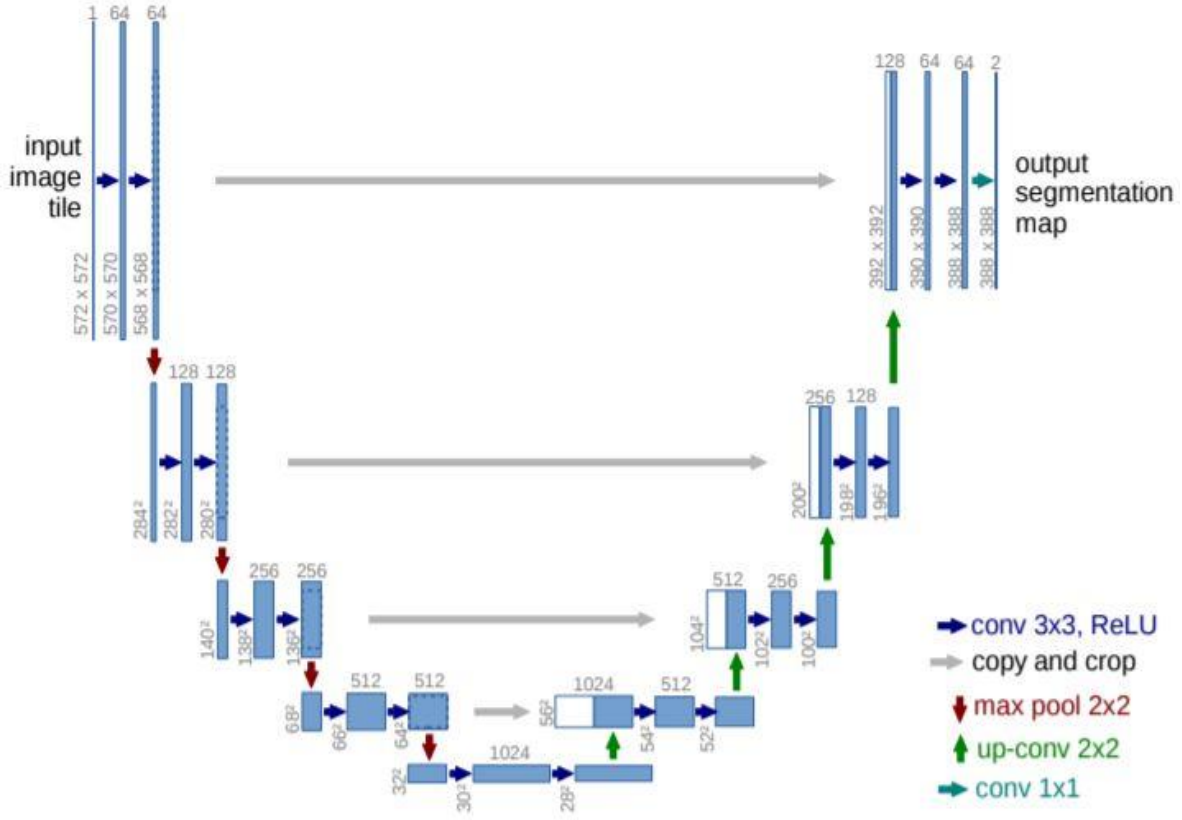

Figure 5-4. The overview of U-Net reprinted from[33].

Ronneberger et al. [33] developed the U-Net architecture for Biomedical Image Segmentation, and this network contains two parts. The first part is known as the encoder or contraction part to encapsulate the context in an image. This part contains several stacks of traditional convolutional layers and max-pooling layers. The second part is the decoder part or symmetric expanding path, which utilized transposed convolutions to enable precise localization. This architecture only includes convolutional layers and does not contain any dense layer so that it can acquire any size of images. Therefore, we can call this network an end-to-end fully convolutional network (FCN).

In the first part of semantic segmentation, we try to predict the left and right lungs for our dataset. We take a total of 704 posterior-anterior (PA) chest radiographs and ground truth (GT) values from Montgomery and Shenzhen chest X-ray (CXR) dataset [49] to train the U-Net architecture. Here, we used 564 CXR images (80 %) for training and 140 (20 %) X-ray images for validation. Then for the testing, we utilize our dataset to predict the segmentation map for both lungs. Specifically, we run the binary segmentation to accomplish this task.

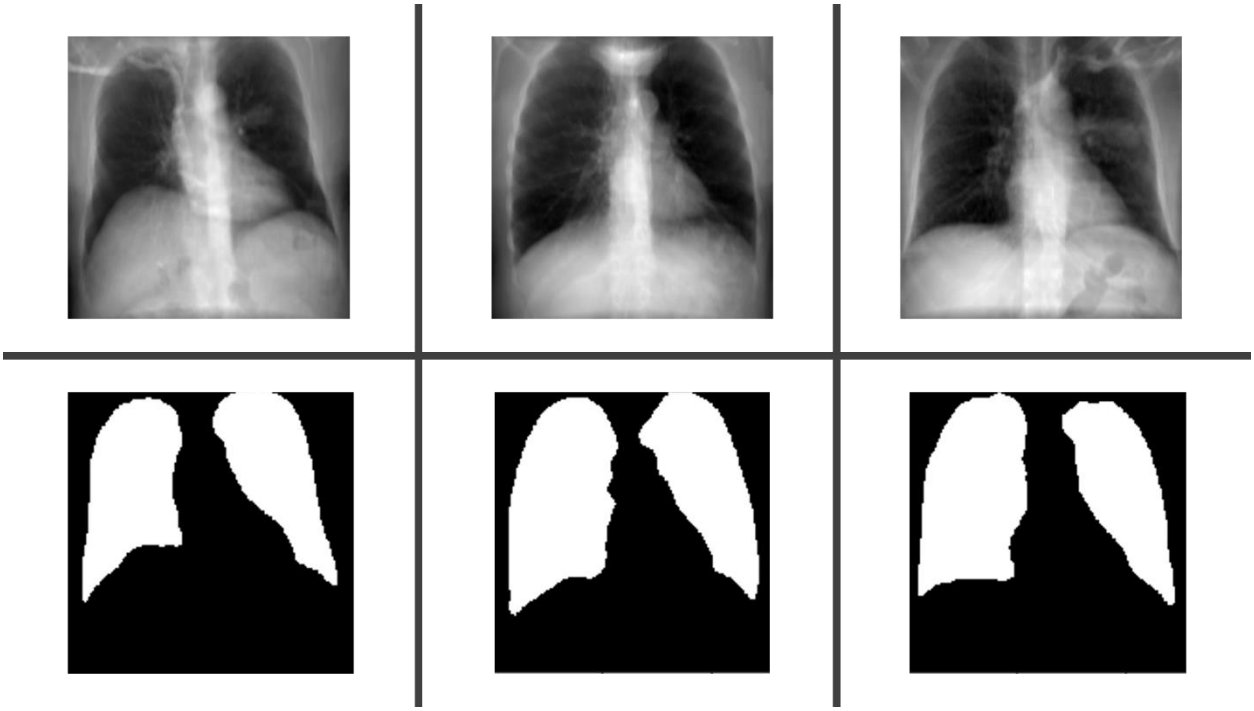

Figure 5-5. Binary segmentation for the lungs using U-Net. The first row is the inputs from our dataset and the second row is the predicted segmentation maps for the corresponding inputs.

Next, to produce the heart segmentation map, we use 247 chest radiographs and the ground truth (GT) value of the heart, respectively, from the JSRT database [50] and SCR database [51]. Furthermore, we split both CXR images and GT into 80 % for training (198 images) and 20 % (49 images) for validation. Then, we execute semantic segmentation on 1018 synthetic X-ray images of our created dataset to predict the 2D segmentation maps.

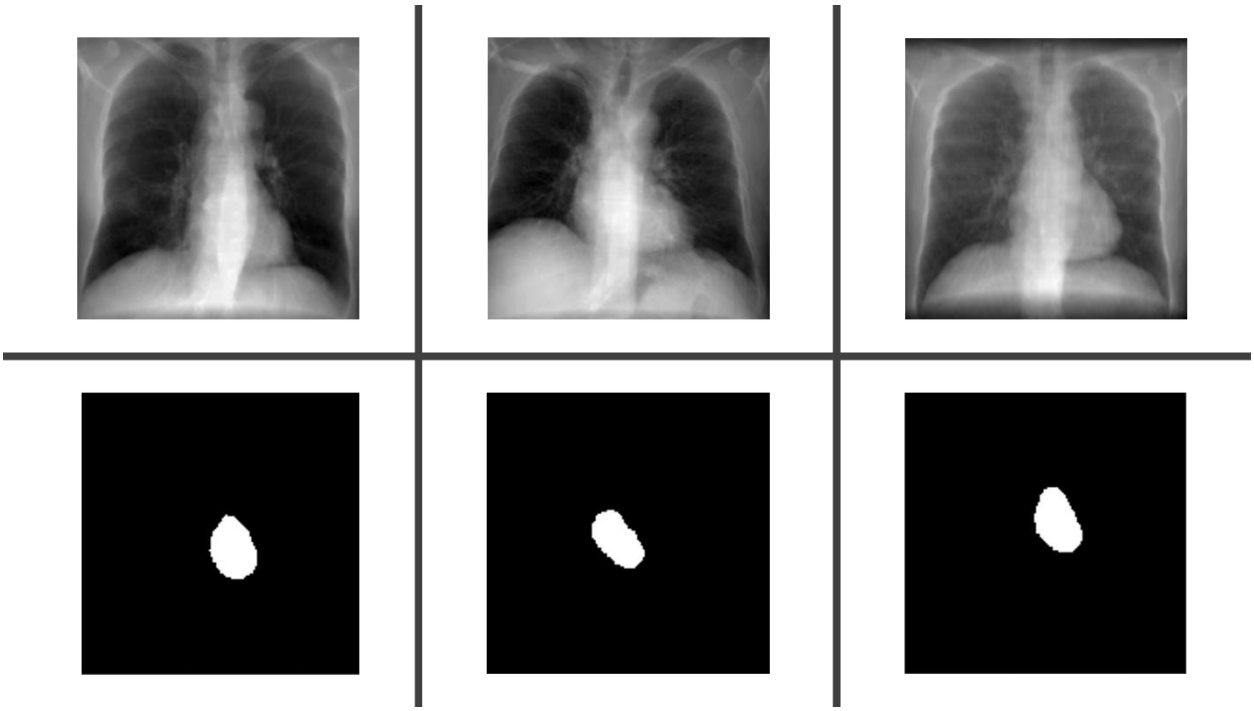

Figure 5-6. Binary segmentation for the heart using U-Net. The first row is the inputs from our dataset and the second row is the predicted segmentation maps for the corresponding inputs.

To lessen the computational cost of the U-Net architectures, we resize and rescale the input image size to 256×256 for all the data. To maintain the same contorted, at first, we crop the center part and then reduce the dimension of the images by the following 1:1 width to height ratio. We inspected different combinations of image shapes (320×320, 512×512, 256×256, 128×128, and 64×64). Therefore, based on space complexity and accuracy, we select 256×256 image dimensions for these two models. For both binary segmentation task, the dimension of input images and produced 2D segmentation map is 256×256.

At last, we combined both binary segmentation task outcomes (lungs and heart) and obtained a multi-categorical segmentation map. Finally, to feed these segmentation maps in our reconstruction model, we resized these maps to 128×128.

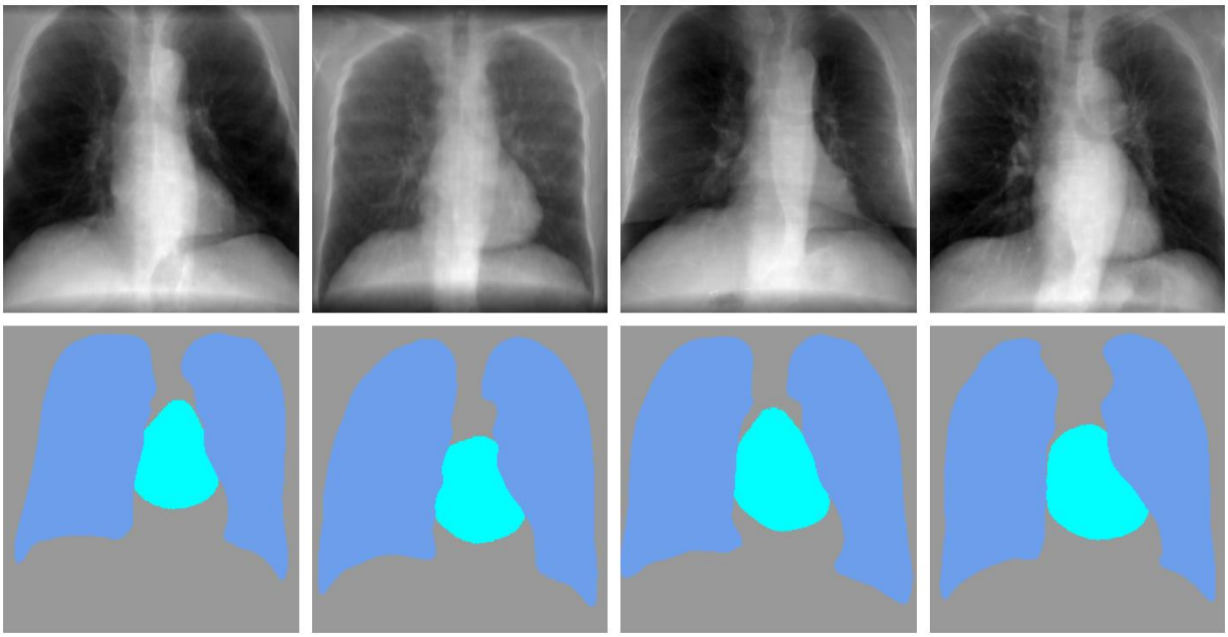

Figure 5-7. Postero-anterior (PA) synthetic chest X-ray images from our dataset (first row) and corresponding predicted 2D segmentation map of lungs and heart from UNet (second row).

5.1.3 **Resized Dataset**

We follow the suggestion of X2CT-GAN to resized the X-ray images to 128×128 for the input of all the networks used in this project. However, Henzler et al. [30] suggested the input image size to 256×256 for their network 2DCNN. At first, we try the approach of 2DCNN for the input image size, though this approach increases the computational complexities in our models, and we do not detect any improvement in the provided 3D volumes from our system.

First of all, to reconstruct the CT from the single view X-ray, we proposed the single view CCX-rayNet where we take the frontal view (posterior-anterior view) X-ray images with 128×128 dimension. Secondly, for the biplanar view of our model, we take a frontal and a lateral view of X-ray images of the same size as the single view network (128×128). Only the frontal X-ray segmentation map required to train the network, and we resized it to the 128×128 dimension. For the segmentation map, we assume two categories, i.e., lungs and heart. A "background" category is used to encompass regions that do not appear in the categories mentioned above. Finally, $128 \times 128 \times 128$ voxels size is specified as the output dimension of all the models.

5.1.4 **Data Standardization**

Data standardization, also known as z-normalization, is one type of normalization technique. The z score denotes the standard score, which is a transformed value of each data point. Data standardization is a vital step in our data pre-processing work because it ensures that each parameter of our dataset (pixel) has a similar data distribution. The standardization process makes convergence faster when we train our network and minimize the computational complexities during the training.

Standardization is a popular data scaling procedure where it assumes that the distribution of the dataset is Gaussian and shifts the distribution to have a standard deviation of one and a mean of zero.

$$z_i = \frac{x_i - mean(x)}{std(x)} \; i = 1,2,3, \dots, n \qquad \text{Eq. 24}$$

z_i= normalized data point

x_i= individual data point

$mean(x)$= the mean value of the entire dataset

$std(x)$= the standard deviation value of the entire dataset

A set called M has total n numbers, so the mean of that set will be:

$$mean = \frac{1}{n}(\sum_{i=1}^{n} x_i) \qquad \text{Eq. 25}$$

The standard deviation of set M will be:

$$std = \sqrt{\frac{1}{n}(\sum_{i=1}^{n}(x_i - mean)^2)} \quad \text{Eq. 26}$$

For normalization of the X-ray image, we take the range n from $0 - 255$ and discard any pixel value which is out of this range. Next, for the CT volume, we reject any pixel value that is not in the range of $0 - 2500$. We follow the same strategy as X2CT-GAN to adjust the pixel value range.

Chapter 6 **Experiment**

This section describes a detailed explanation of our project setups, training settings, and optimization function. Next, we quantitively evaluate our proposed CCX-rayNet with two widely used evaluation matrics called structural similarity (SSIM) index and peak signal-to-noise ratio (PSNR). Furthermore, we reproduce the baseline methods named X2CT-GAN and X2CT-CNN to manifest the efficacy of our proposed system. In order to demonstrate the credibility of our network, we provide a detailed qualitative analysis based on the produced 3D volume from our model and then compare the visualization result with the state-of-the-art method (X2CT-GAN). Lastly, we perform an ablation study to indicate the importance of every proposed module. We exhibit the outcome for different setups in the ablation study section.

6.1 **Setups**

6.1.1 **Libraries and Application**

We used three different setups to accomplish our whole project.

1. The first setup has 7 core CPU with 16 GB of memory facilitated with an Nvidia GPU GTX1080.
2. The second setup has 9 core CPU with 32 GB memory equipped with an Nvidia GPU RTX2080Ti.
3. The third setup for this project has 7 core CPU with 32 GB of memory facilitated with two Nvidia GPU RTX2080Ti.

We used python programming language and several deep learning libraries of python, such as Keras 2.3.1, Tensorflow 2.1.0, and Pytorch 1.3.0, to accomplish the whole Deep Learning part of this project. Furthermore, for data pre-processing and dataset making, we utilize PyDicom and

OpenCV libraries. Finally, to visualize the 3D outcome and inspect the result, we use the 3D slicer and ITK-SNAP software.

6.1.2 Training Setting for U-net

For the segmentation map, we train both U-Net with similar training settings. Both the networks train for 400 epochs or iterations with a batch size of 16. We employ the Adam optimization function with a learning rate $1e-5$ for both of the training. Moreover, momentum parameters are $\beta_1 = 0.9$ and $\beta_2 = 0.999$ same as default and epsilon is none. Moreover, for the training of U-Net, we use the dice coefficient loss function. We utilize two callback functions for the training part of U-Net to complete some routine work.

Firstly, the method called EarlyStopping applies to inspect whether validation loss has been decreased or not for 15 epochs. If the validation loss remains identical after 15 epochs of training, then we stop the training and run the test process. Otherwise, it will run until complete the whole 400 epochs of training.

The second callback function, namely, ReduceLROnPlateau, monitors the learning rate during the training period. When the validation loss is not diminishing for seven epochs, we lessen the learning rate as the factor by $\sqrt{0.1}$. So, the new learning rate x_{new} :

$$x_{new} = x * \sqrt{0.1} \qquad \text{Eq. 27}$$

Here, x_{new}= new learning rate; x = current learning rate. Furthermore, the lower bound of the learning rate is $0.5e-6$.

6.1.3 Training Setting for CCX-rayNet

We train the CCX-rayNet for the 100 epochs with a batch size of 2 because of GPU memory limitations. We use Adam as the optimizer with a learning rate initialized as $2e-3$. Moreover, momentum parameters are $\beta_1 = 0.5$ and $\beta_2 = 0.999$, and epsilon is none. Additionally, we decay

the initial learning after 50 epochs with 30 percent. Besides, We use instance normalization instead of batch normalization as instance normalization proves to be superior in image generation task.

6.1.4 Adam Optimizer

Adam optimization function [53] is designed for deep neural networks, specifically for adaptive learning rate optimization. This algorithm is based on the adaptive approximation of lower-order moments, which particularly for the first-order gradient-based optimization process of stochastic objective functions. This method can be described as the combination of Stochastic Gradient Descent (SGD) and Root Mean Square Propagation (RMSProp) with momentum. Like RMSProp, it scales the learning rate by using squared gradients, and instead of the gradient itself, it uses the moving average of the gradient with momentum like Stochastic Gradient Descent (SGD) with momentum. Furthermore, Stochastic gradient descent (SGD) preserves the same for all weight updates, and it does not change during the training. On the other hand, Adam computes the individual adaptive learning rate for each weight (parameters) of the network and independently adapted as learning unfolds. Conclusively, this method is computationally efficient, simple to implement, demand a small amount of memory, and, most importantly, well match for the large dataset and/or huge parameters.

6.2 Evaluation Metrics

6.2.1 PSNR

The peak signal-to-noise ratio is also known as PSNR, is used to calculate the ratio between the highest obtainable power of a signal and the power of corrupting noise that influences the constancy of its representation. PSNR is usually used to yield the difference between two signals. Moreover, it can be normalized by the signal extent and generally exhibited in respect of the logarithmic decibel scale. Through the PSNR, we try to measure the density recovery of our produced CT volume. From the PSNR value, we can find the similarity between the original and reconstructed CT volumes and measure the effectiveness of the proposed algorithms.

$$PSNR = 20\log_{10}(\frac{Max_f}{\sqrt{MSE}}) \quad \text{Eq. 28}$$

Here, Max_f denotes the maximum signal value in our original (ground truth) 3D CT scan, and MSE means the Mean Squared Error (MSE) value. MSE grants us to compare the "true" pixel values of our original CT scan to our reconstructed CT volume.

6.2.2 **SSIM**

The Structural Similarity Index (SSIM) [54] is one of the key criteria of image reconstruction task to evaluate the image quality degradation produced by losses in data transmission or data compression. As opposed to the Peak signal-to-noise ratio (PSNR), SSIM relies on the visible structures of the images. SSIM estimates the resemblance between a reference image and a processed image by considering structure information, brightness, and contrast. In addition, this metric also incorporates contrast masking terms and luminance masking. It is broadly utilized to compute the image reconstruction quality and in many image transfer applications. In our work, we try to evaluate the predicated CT volume and the corresponding ground-truth value through the SSIM metric in order to inspect the resemblance between the two 3D CT volumes.

In Eq. 29, we compute the resemblance between two different 3D CT volume. The operation can be formulated as:

$$SSIM(m,n) = \frac{(2\mu_m\mu_n + c_1)(2\sigma_{mn} + c_2)}{({\mu_m}^2 + {\mu_n}^2 + c_1)({\sigma_m}^2 + {\sigma_n}^2 + c_2)} \quad \text{Eq. 29}$$

Where we assume that m, n are two different 3D CT volume (m: generated 3D CT volume, n: ground-truth value) with a common size $P \times P$. Furthermore, the parameters of Eq. 29 can be defined as:

- The average of m is μ_m
- The average of n is μ_n
- The variance of m is ${\sigma_m}^2$
- The variance of n is ${\sigma_n}^2$

- The covariance of m, n is σ_{mn}
- The dynamic range of the pixel values is L
- Two weak denominators to stabilize the division: $c_1 = (J_1 L)^2$, $c_2 = (J_2 L)^2$
- By default, $J_1 = 0.01$ and $J_2 = 0.03$

The SSIM measured on three comparison measurements between m, n: Luminance (l), structure (h), and contrast (c).

$$l(m,n) = \frac{(2\mu_m \mu_n + c_1)}{({\mu_m}^2 + {\mu_n}^2 + c_1)} \quad \text{Eq. 30}$$

$$h(m,n) = \frac{\sigma_{mn} + c_3}{\sigma_m \sigma_n + c_3} \quad \text{Eq. 31}$$

$$c(m,n) = \frac{2\sigma_m \sigma_n + c_2}{{\sigma_m}^2 + {\sigma_n}^2 + c_2} \quad \text{Eq. 32}$$

Here, $c_3 = \frac{c_2}{2}$

6.3 Qualitative Analysis

CT reconstruction from X-ray images is a newly proposed task, and previously only a few research papers were published based on these chores. Furthermore, here we try to reconstruct the CT volumes from thorax or chest X-ray (CXR) image, where we need to consider many significant issues because many vital organs of the human body existed between the neck and abdomen. For instance, the heart and lungs are two major organs in this area with many complex internal structures. Additionally, many veins and blood vessels, trachea, larynx, and some parts of the sternum are also included in this part of the body. Besides, we have to intuitively inspect several essential bone structures such as the ribs cage, a specific part of the cervix or neck, clavicles, and vertebral column or backbone to reconstruct the CT scan. Therefore, we have to try several

different setups and carefully inspect every model's output to detect fine details close to a ground truth value.

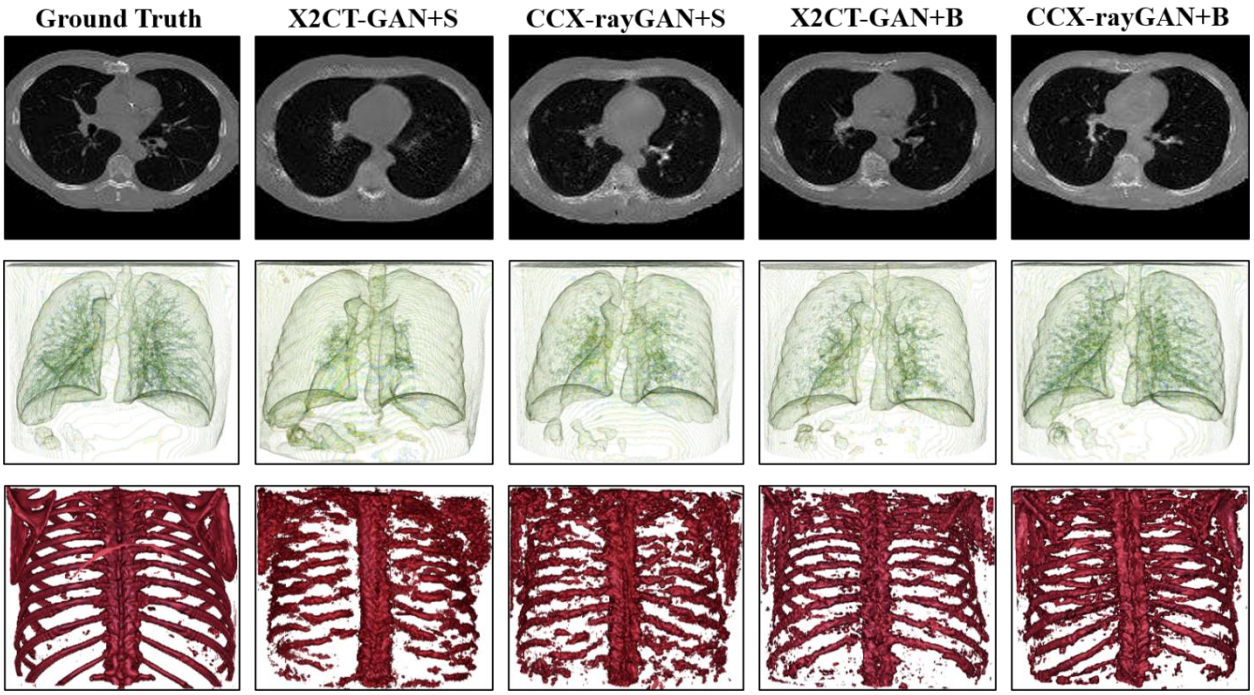

Figure 6-1. 3D CT volumes (posterior-anterior view) are reconstructed from several models. CCX-rayGAN is our proposed method, and X2CT-GAN [2] is the baseline approach. Here, '+S' and '+B' denotes single view X-ray image input and biplanar X-ray image input, respectively. The first row exhibits the axial view from the reconstructed CT volume, the second row indicates the posterior-anterior view (PA) view of the CT scan, and the third row signifies the bone structure reconstruction by several networks.

Recently GAN has been hugely implemented in the medical domain and helps researchers to achieve the result with critical anatomical details. Therefore, here we proposed a GAN-based model, namely CCX-rayGAN, to improve the perceptual quality of the reconstruction result. At first, we implement the single view GAN-based method, and then to enhance the visual quality near to ground truth CT value, we build the biplanar view architecture with GAN like setup.

For a fair comparison, we reproduce both GAN based single view and biplanar view architecture of X2CT-GAN. X2CT-GAN is the only method until this date, which performs the same task as ours (CXR image to CT scan) by using the LIDC-IDRI dataset. We qualitatively compare X2CT-GAN with our methods and the outcome displayed in Figure 6-1. As shown in Figure 6-1, '+S' means single view X-ray image input, and '+B' means biplanar X-ray image input. We provide the output in three different ways to compare our results with the state-of-the-art methods. In the

first row of figure 6-1, we exhibit the axial view from the reconstructed CT volume. The second row indicates the posterior-anterior (PA) view of the CT scan. In this view, we try to compare the result based on both lungs, all the critical veins inside the lungs (pulmonary veins and arteries), trachea, and larynx. In the third or the final row, we try to demonstrate the bone reconstruction result of our models. Here we show the visual comparison of several bone structures such as the ribs cage, both clavicles, scapulas, and the vertebral column.

From the first and second row of Figure 6-1, we can evaluate the visual quality from two different viewpoints, and we can easily see the differences and improvements in our proposed architecture. First of all, CCX-rayGAN+S generates sharper boundaries of the large organs, and the internal textures of the organs are well reconstructed as well. Furthermore, compared to X2CT-GAN+S, our single view GAN system can produce a more detailed and accurate result, which we can see from the first two rows. The model successfully reproduces more intricate details than the baseline single view system. Secondly, the reconstruction results of our proposed biplanar view architecture, namely CCX-rayGAN+B, can yield the visual output, which is more close to the ground truth values. Most importantly, the axial view and posterior-anterior (PA) view of the CT scan indicate the superiority of the biplanar mode of CCX-rayGAN in order to recover the outline of the body and maintains small complex anatomical structures such as pulmonary veins and arteries (small vessels inside the lungs). Moreover, CCX-rayGAN+B outperforms the X2CT-GAN+B on reconstructing the shape and texture of the important anatomies. The improvement can be easily seen from the visualization results, and the CCX-rayGAN+B also produces accurate boundaries of lungs than the baseline X2CT-GAN model.

Finally, in the third row of Figure 6-1, we show the backside's bone structure (area near the backbone). We inspect some main bone structures like the rib cage, clavicles, scapulas, and vertebral column to evaluate bone reconstruction. Our proposed biplanar model (CCX-rayNet+B) overall exhibits superior outcomes than the state-of-the-art model (X2CT-GAN+B). Especially for the rib cage, vertebral column, and scapulas, the model shows a precise reconstruction accuracy. Furthermore, the bone reconstruction result of CCX-rayNet+B is closer to the ground truth than the X2CT-GAN+B.

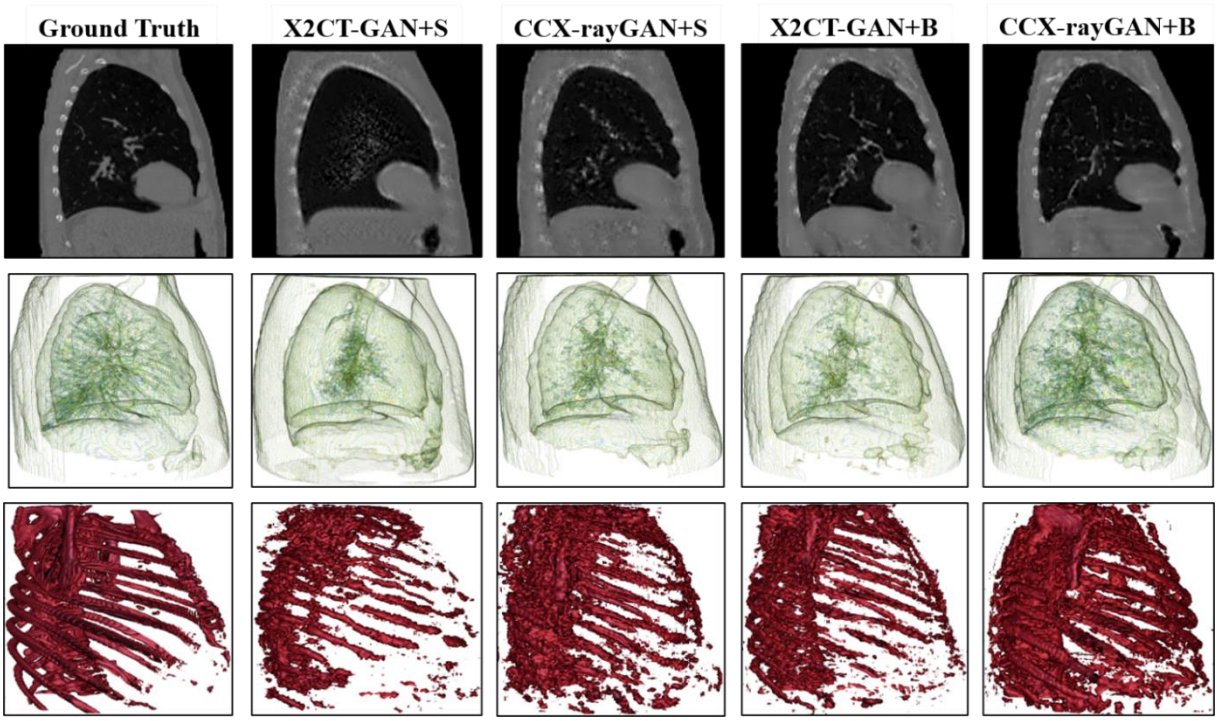

Figure 6-2. 3D CT volumes (lateral view) are reconstructed from several models. CCX-rayGAN is our proposed method, and X2CT-GAN [2] is the baseline approach. Here, '+S' and '+B' denotes single view X-ray image input and biplanar X-ray image input, respectively. The first row exhibits the sagittal view from the reconstructed CT volume, the second row indicates the lateral view of the CT scan, and the third row signifies the bone structure reconstruction by several networks.

In Figure 6-2, we show the lateral view of 3D volume reconstruction to inspect the result critically. Same as Figure 6-1, we can find that our biplanar view of CCX-rayGAN displayed a superior outcome with more detailed results. Moreover, we can spot the small vessels inside the lungs very clearly, which is better than X2CT-GAN+B. Similarly, our proposed CCX-rayGAN+S outperforms X2CT-GAN+S based on the reconstructed output.

We evaluate the visual quality of our prediction from Figure 6-3, and it is obvious to see the differences and improvements. First of all, CCX-rayGAN+S generates sharper boundaries of the large organs compared to the X2CT-GAN+S, and the internal textures of the organs are well reconstructed as well. Moreover, through the lateral and axial view of a 3D CT volume, we demonstrate that our suggested CCX-rayGAN+S can yield high-quality reconstruction results for intricate anatomical parts.

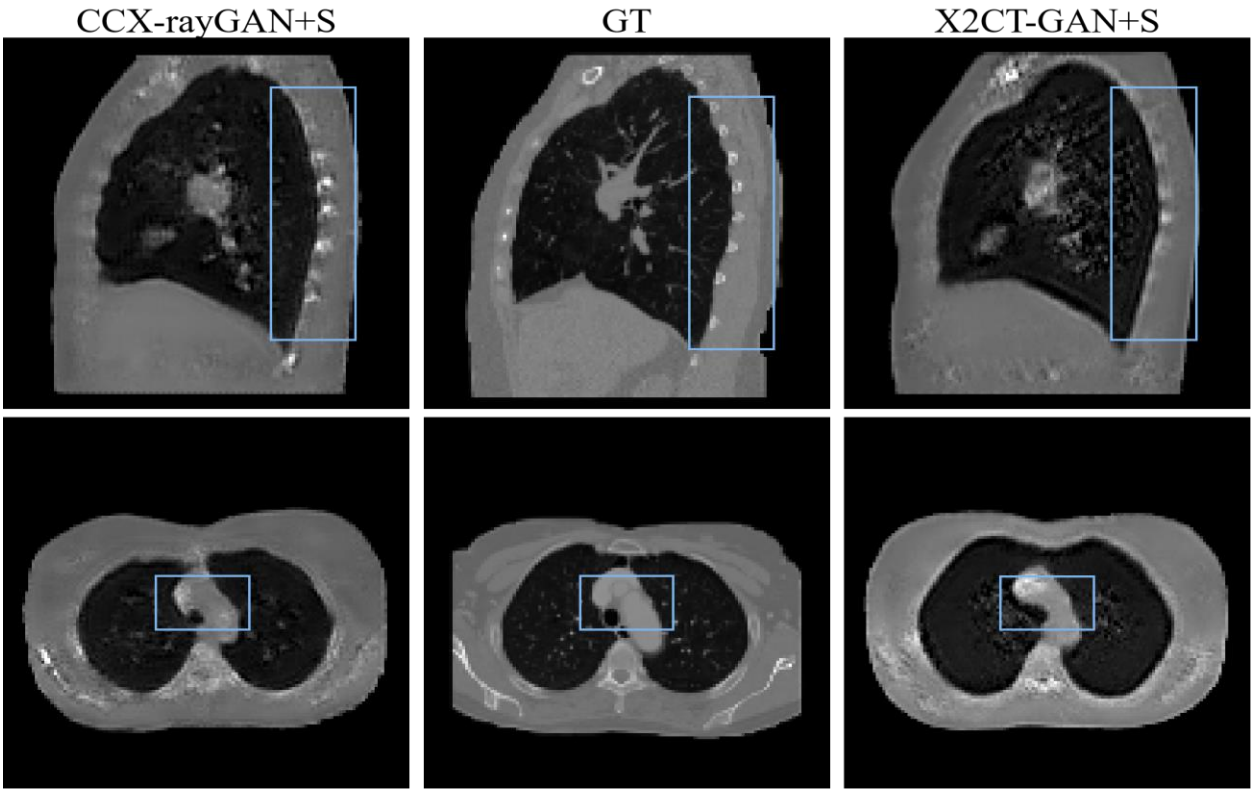

Figure 6-3. Lateral view (first row) and axial view (second row) from a CT volume yielded by CCX-rayGAN+S in the first column, single view X2CT-GAN+S [2] in the third column, and ground truth displayed in the second column. Our proposed single view model provides precise anatomical reconstructions than X2CT-GAN.

In Figure 6-4, we can easily see that our biplanar view method also yields superior outcomes and outperforms the biplanar view of X2CT-GAN. It is obvious to see the intricate details from the output of our biplanar view method, which is visually more close to ground truth value than the baseline method. Additionally, the biplanar view remarkably enhances the quality of the output. We can observe this significant improvement in our outcome through Figures 6-1, 6-2, 6-3, 6-4. Thus, our architecture can achieve more complex anatomical structures by biplanar view setting, which is not possible through the single view model.

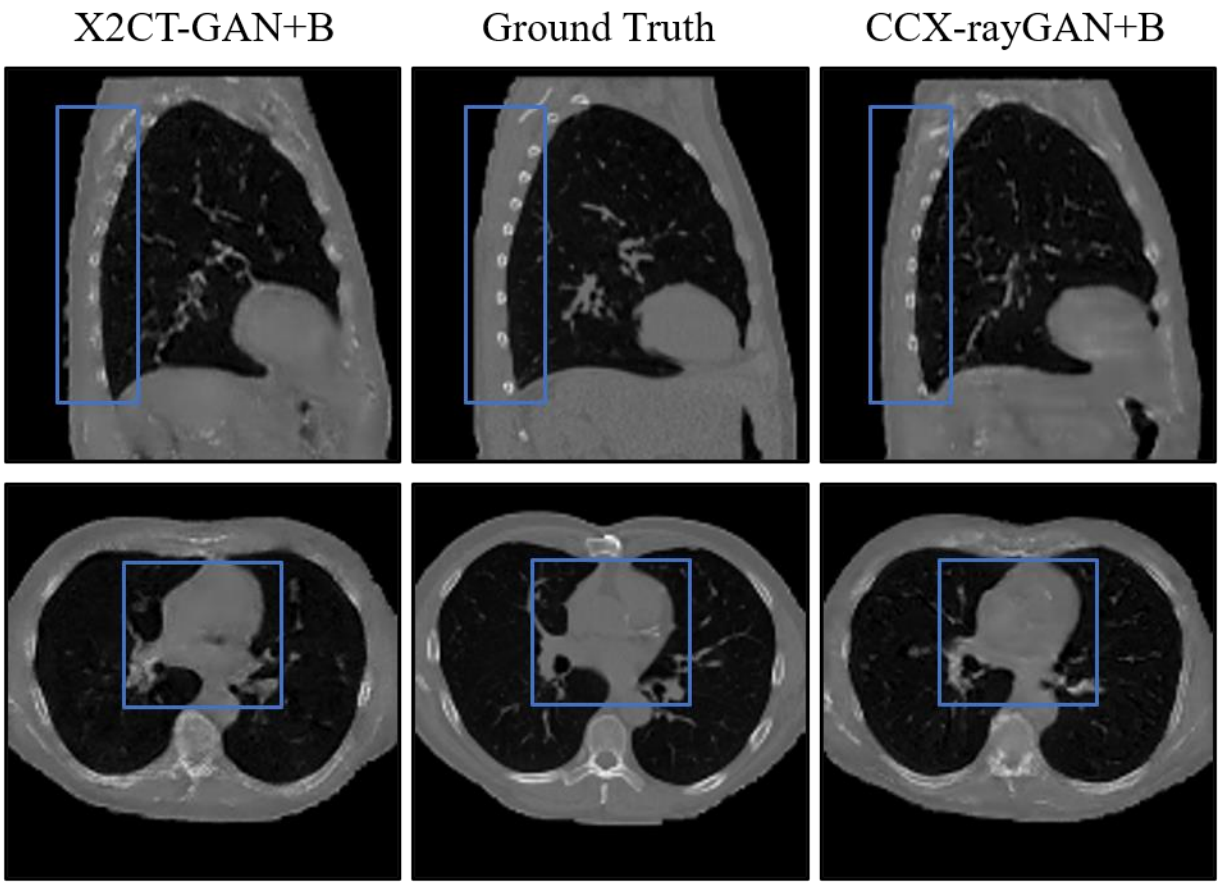

Figure 6-4. Lateral view (first row) and axial view (second row) from a CT volume yielded by CCX-rayGAN+B in the third column, single view X2CT-GAN+B [2] in the first column, and ground truth displayed in the second column. Our proposed biplanar view model provides precise anatomical reconstructions than X2CT-GAN+B.

6.4 **Quantitative Analysis**

As shown in Table 3, our proposed single view model, namely 'CCX-rayNet+S' generates superior PSNR and SSIM accuracy. This method surpasses all the popular single view reconstruction model. In addition, CCX-rayNet+S achieves more than 1.6 dB PSNR value than the 'X2CT-CNN+S'. Additionally, the state-of-the-art single view model (X2CTCNN+S) provide a 0.587 SSIM value, whereas our proposed single view model exhibits 0.597 SSIM, which is higher among the single view models.

Table 3. Quantitative results from proposed CCX-rayNet, X2CT models [2], and 2DCNN [30]. CCX-rayNet is the proposed generator network, and CCX-rayGAN' is our GAN based method. Here, '+S' denotes the single-view X-ray image input into the CCX-rayNet. Besides, the best results are highlighted in **bold.**

Model	PSNR (dB)	SSIM
2DCNN	23.10	0.461
X2CT-GAN+S	22.30	0.525
X2CTCNN+S	23.12	0.587
CCX-rayGAN+S	22.48	0.536
CCX-rayNet+S	**24.73**	**0.597**

According to Table 4, the proposed biplanar CCX-rayNet provides superior PSNR and SSIM values than the state-of-the-art method (X2CT-CNN+B). We see a similar trend for the SSIM. For instance, CCX-rayNet+B achieves exactly 3.1% improvement than the X2CT-CNN+B. So, in both evaluation metrics, our proposed CCX-rayNet+B present superiority over the baseline methods.

Table 4. Quantitative results from proposed CCX-rayNet, X2CT models [2], and 2DCNN [30]. CCX-rayNet is the proposed generator network, and 'CCX-rayGAN' is our GAN based method. Here, '+B' indicates the biplanar X-ray input into the CCX-rayNet. Besides, the best results are highlighted in **bold**.

Model	PSNR (dB)	SSIM
X2CT-GAN+B	26.19	0.626
X2CTCNN+B	27.29	0.721
CCX-rayGAN+B	26.93	0.694
CCX-rayNet+B	**28.18**	**0.752**

The biplanar training process significantly enhances the reconstruction performance for both CCX-rayNet+B and CCX-rayGAN+B models than their single view counterparts. Our biplanar method (CCX-rayNet+B) almost enhances 3.4 dB than our single view CCX-rayNet. Moreover, the biplanar network obtains 15.5% more SSIM value than the single view model.

6.5 Ablation Study

Table 5. Evaluation of several settings of proposed CCX-rayNet+S and the best outcome marked in **bold**. Here, we display the experiment analysis for the single-mode ('+S').

DFT	×	✓		✓
DAC	×		✓	✓
PSNR (dB)	23.12	24.26	23.91	**24.73**

In this part, we conduct an ablation study to assess the potency of different proposed modules in our CCX-rayNet. The AFF module is exclusively used in the biplanar view model. In addition single view contains DFT and DAC modules. Thus, the study is divided into two parts, single view, and biplanar view shown in Table 5 and Table 6, respectively. We mainly focus on these factors:

1. The combination of deep feature transform (DFT) module
2. depth-aware connection (DAC), and
3. The adaptive feature fusion (AFF) structure.

As shown in Table 5, both DFT and DAC give a decent improvement individually for the single view CCX-rayNet. Furthermore, the PSNR value has been elevated to almost 0.82 dB when we incorporate both DFT and DAC modules together. The result shows the effectiveness of these two modules inside the CCX-rayNet+S.

Subsequently, from Table 6, we can pronounce that the proposed biplanar CCX-rayNet with a combination of DFT, DAC, and AFF modules can achieve a superior PSNR accuracy. This

indicates that all three modules play a vital role in reinforcing the network performance for the biplanar version, and among them, DFT makes the most contribution.

Table 6. Evaluation of several combinations of proposed CCX-rayNet+B and the best outcome marked in **bold**. Here, we show the experiment analysis for the biplanar mode ('+B').

DFT	×	✓	✓	✓
DAC	×		✓	✓
AFF	×			✓
PSNR	27.29	27.68	27.49	**28.18**

6.6 **Result Analysis**

Our proposed GAN networks provide significantly better visual image quality (in Figure 6-1, 6-2, 6-3, and 6-4) with subtle details while sacrificing a small amount of PSNR value, which is usual in any GAN based architectures. The GAN based architectures habitually relinquish MSE-based metrics to attain better visualization output. A similar phenomenon we also discover in this work. Therefore, we can draw up a logical trade-off in this matter. For example, our CCX-rayNet+B only gain 1.25 dB more PSNR value than our biplanar GAN based method, while the GAN based approach dramatically enhances the visual standard. From figure 6-5, we can easily detect this tendency of our networks. GAN based architecture precisely maintains many complicated anatomical structures, whereas CCX-rayNet+B provides very blurry 3D volumes. However, CCX-rayNet+B can generate sharper boundaries around large organs such as the chest wall, lungs, and heart. We can assert that the quality of visualization is valuable for any medical professional to observe the necessary details. Therefore, in the medical domain, it is beneficial to produce images with complex anatomical structures.

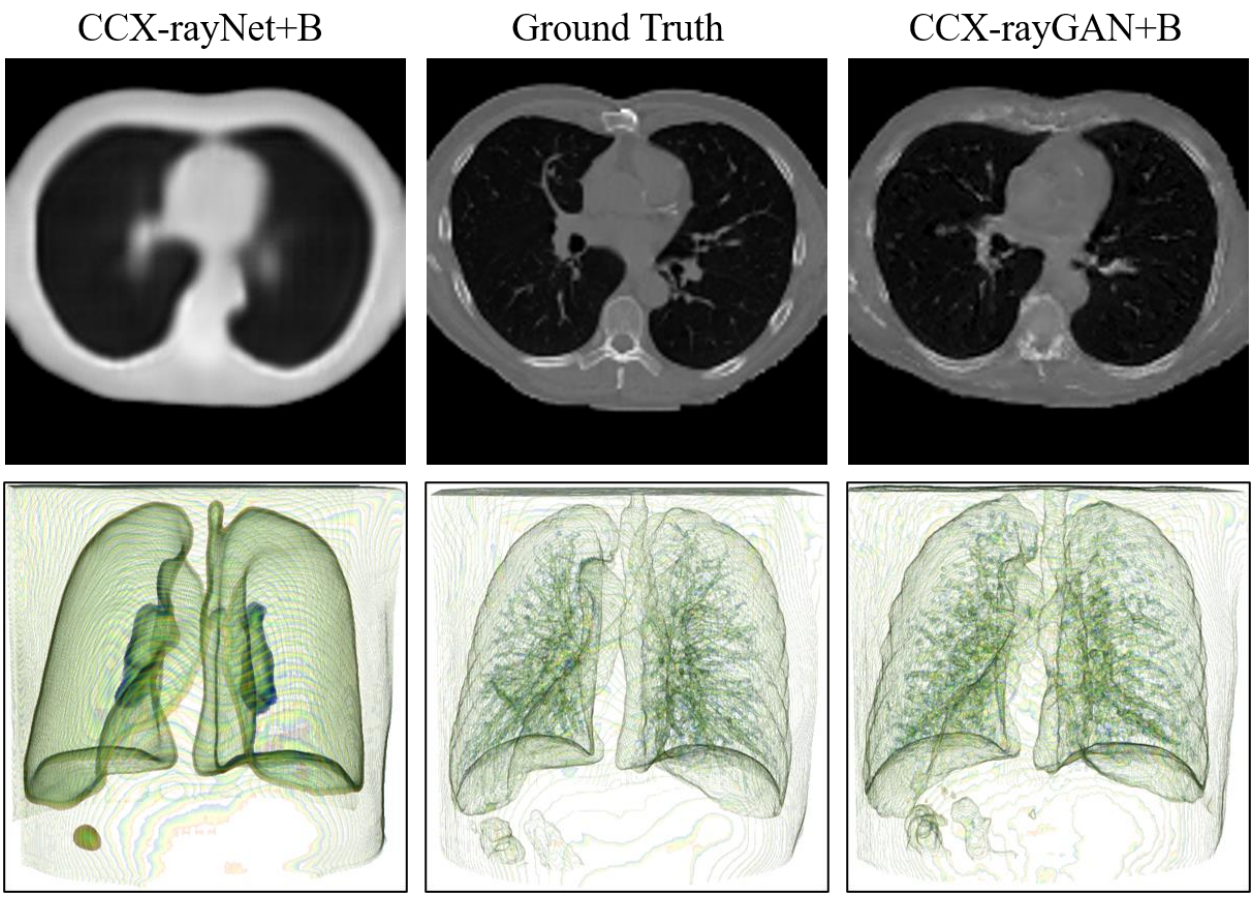

Figure 6-5. The first row exhibits the axial view from the reconstructed CT volume, the second row indicates the posterior-anterior (PA) view of the CT scan, and '+B' means the biplanar view. We exhibit that our proposed GAN based biplanar view method produces higher visual quality than our proposed CCX-rayNet+B.

We also achieve the same characteristics from our single view proposed model. From Figure 6-6, we can discover that CCX-rayNet+S can provide sharper boundaries around the organs, but it can not generates complex details like CCX-rayGAN+S.

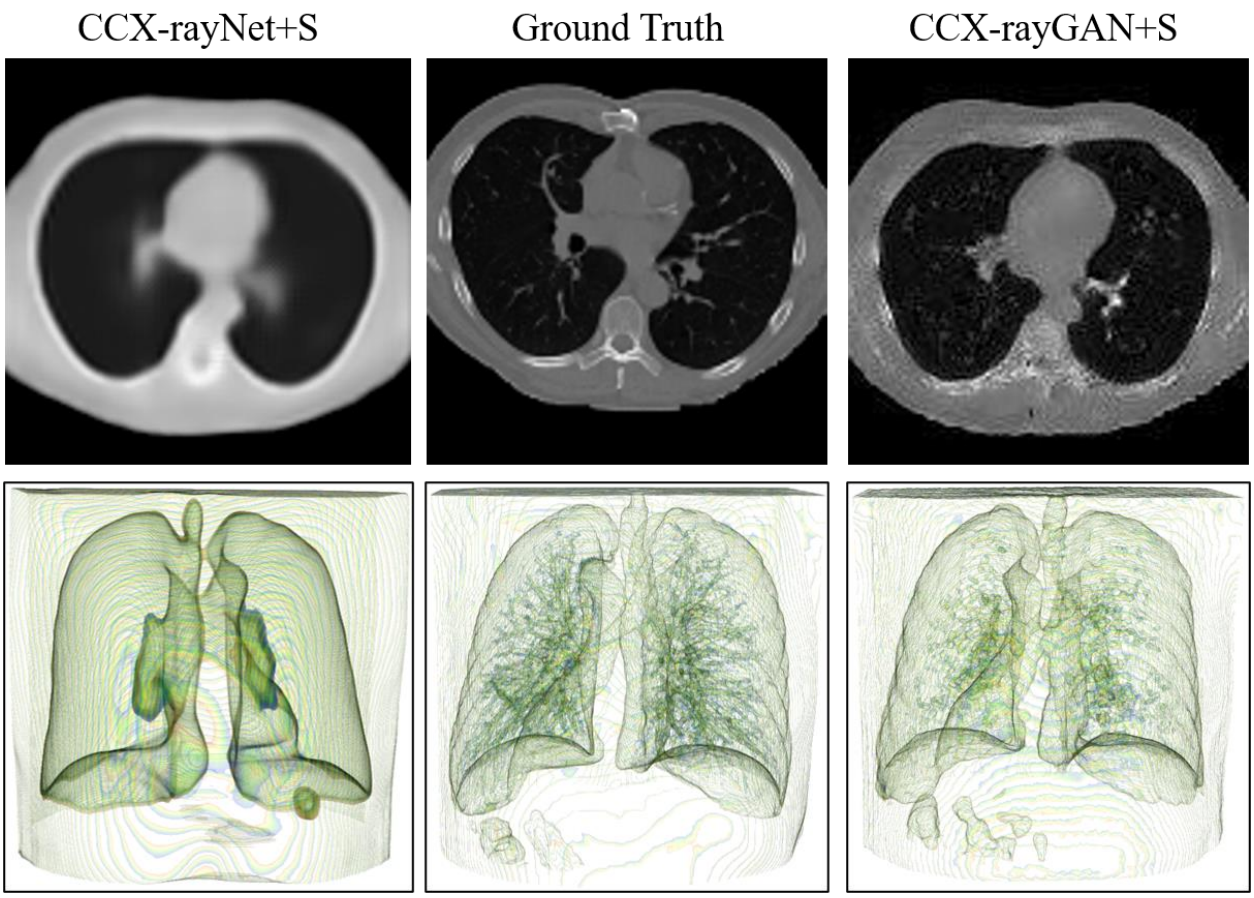

Figure 6-6. The first row exhibits the axial view from the reconstructed CT volume, the second row indicates the posterior-anterior (PA) view of the CT scan, and '+S' means the single view. We exhibit that our proposed GAN based single view method produces higher visual quality than our proposed CCX-rayNet+S.

We tried several other models to reconstruct the 3D CT volumes:

1. Based on channel-wise attention mechanisms [58]: 2D channel-wise attention based encoder and 3D channel-wise attention based decoder
2. U-Net lookalike a CNN based 3D reconstruction model

However, we find ambiguities in the results on U-Net like architecture, which is not complicated enough to produce the 3D CT volumes. Attention-based architecture can generate the 3D volume, although the product is not satisfactory enough. Moreover, the channel-wise attention mechanism enhances the computational complexities in our networks.

Tackling the architectural complexities was one of the main challenges in this entire thesis work. We have to build a complicated model to produce a high-quality 3D CT volume. For example, our biplanar CCX-rayGAN model comprises 101.3 million parameters, which takes a long time to train using a powerful setup. So, we have to draw up a trade-off between excellent visual quality output and the architectural complexities.

Chapter 7 **Conclusion**

The X-ray image to CT scan reconstruction utilizing the deep neural networks (DNN) is considerably a new idea in the medical field. To be specific, by using DNN, this is the second work to reconstruct the CT volume from chest X-ray (CXR) image as far as we can aware. This thesis presented a new class-conditioned deep neural network, namely CCX-rayNet, to reconstruct 3D CT volume from synthesized chest X-ray images. CCX-rayNet is a deep convolutional neural network (DCNN) to assist the medical domain. Next, in order to achieve small details and accurate anatomical structure of some major organs such as lungs, heart, liver, rib cage, we apply a generative adversarial network (GAN) with CCX-rayNet. In this work, we proposed two separate ways to accomplish the CT volume. Firstly, the single-mode approach where the inputs are posterior-anterior view X-ray images and the semantic segmentation map of the frontal view of X-ray images. Secondly, the biplanar mode approach where we take the posterior-anterior view and lateral view X-ray images with the semantic segmentation map of the frontal view of X-ray image as the inputs of the proposed method. To recovering textures and shapes faithful to semantic classes in the real-world scenario, we proposed a deep feature transform (DFT) module, depth aware connection, and an adaptive feature fusion (AFF) module. Experimental analysis exhibit that, in the 3D reconstruction process, our proposed biplanar model with prior semantic knowledge reinstates the internal anatomical structure, density information, and shape with a higher standard than any other state-of-the-art methods. Our significant contributions and some likely applications ideas are following.

7.1 **Accomplishment**

We propose a new GAN network for the medical domain, namely CCX-rayGAN, which can generate a fine quality 3D CT volume after dealing with the issues of distinct modalities and dimensionalities in the input and output (X-ray image to CT), whereas the previous methods only tackling the case of the same dimensionalities. Simultaneously, we required at most two X-ray

images as input to produce the output; however, the previous techniques required hundreds of X-ray images to accurately reconstruct a CT volume. In addition, we utilize semantic prior information from the segmentation map to train the network. Our network can achieve valuable information about different organs from these segmentation maps.

7.2 Major Contribution

The contributions of our work are three folds:

- **Semantic Prior Information with DFT:** As far as we know, this is the first work that exerts the semantic prior information to reconstruct the 3D CT volume from the 2D chest X-ray (CXR) image. Categorical priors are valuable to reconstruct the chest CT volume because, in the chest, the different organs have different semantics (e.g., heart, lungs, liver, rib, spine, etc.). Specifically, the internal structure of the lungs and heart area have distinct textures and should be treated differently. So, we proposed a deep feature transform (DFT) module that modulates the 2D feature maps conditioned on the semantic information spatially. Thus, the proposed CCX-rayGAN can adaptively generate local visual patterns with critical internal anatomical structures.
- **Depth Aware Connection (DAC):** To prevent the depth information loss during the bridging of 2D and 3Dfeatures, we proposed a new module called depth aware connection (DAC). DAC bridged the encoded 2D feature and decoded the 3D feature in a depth aware manner so that the connection becomes more realistic and reduce abundant features.
- **Adaptive Feature Fusion (AFF) for Biplanar View:** It is easier to expand the single view network to a biplanar view; however, there can be a huge possibility that the model will face a registration problem because, in the previous method, average sum operation is

adopted to fuse the features from orthogonal views where the movement of the patient ignored during the biplanar X-ray images capturing. Therefore, to minimize this issue, we replace the average sum operation with a weighted sum operation to bypass the less contributed feature. Specifically, the adaptive feature fusion (AFF) module is based on the weighted sum to fuse multiple views of features. AFF module suppresses the unregistered features and amplifies the most effective ones by employing the attention mechanism. Besides, our proposed biplanar view models achieve superior qualitative and quantitative outcomes than the proposed single view architectures, which ensure that we gain more information through this approach.

7.3 Practical Life Applications

This work's motivation is not to substitute the CT scan apparatus, although this proposed architecture can produce some useful clinical practices appliances. Firstly, we can estimate the size and shape of major organs such as the liver, heart, lungs through this method. Secondly, through this method, the physicians can inspect the CT scan like 3D CT volume with particular clinical values so that we can use this method as an extended version of an X-ray machine. Furthermore, we can utilize this model for diagnosing ill-positioned organs with the help of reconstructed CT volume. In many developing countries, the CT scan machine is still a less accessible appliance. Thus, our method can apply those places with the X-ray machine to assist physicians in examining patients. Finally, in radiation therapy, we can apply this for dose planning.

7.4 Future Work

Our method can reconstruct high-quality CT volumes with detailed anatomical structures and shapes; however, we can inspect small anatomies of the chest that still suffer from little artifacts. So, in the future, we want to resolve the issue to provide the medical community with more precise results. As we know that chest reconstruction is a difficult task due to the presence of many critical

anatomical structures and vital organs. Hence, to improve the visual quality, we need more CT and X-ray images paired data to train our model. From now on, we want to reconstruct more precise CT volumes with perfect shapes and details with the assistance of physicians to fulfill the clinical values.

Chapter 8 **References**

[1] Fred, H. L. (2004). Drawbacks and limitations of computed tomography: views from a medical educator. Texas Heart Institute Journal, 31(4), 345.

[2] Ying, X., Guo, H., Ma, K., Wu, J., Weng, Z., & Zheng, Y. (2019). X2CT-GAN: reconstructing CT from biplanar X-rays with generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 10619-10628).

[3] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).

[4] Herman, G. T. (2009). Fundamentals of computerized tomography: image reconstruction from projections. Springer Science & Business Media.

[5] Wang, X., Yu, K., Dong, C., & Change Loy, C. (2018). Recovering realistic texture in image super-resolution by deep spatial feature transform. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 606-615).

[6] Drebin, R. A., Carpenter, L., & Hanrahan, P. (1988). Volume rendering. ACM Siggraph Computer Graphics, 22(4), 65-74.

[7] Hounsfield, G. N. (1973). Computerized transverse axial scanning (tomography): Part 1. Description of system. The British journal of radiology, 46(552), 1016-1022.

[8] Lustig, M., Donoho, D., & Pauly, J. M. (2007). Sparse MRI: The application of compressed sensing for rapid MR imaging. Magnetic resonance in medicine: An official journal of the international society for magnetic resonance in medicine, 58(6), 1182-1195.

[9] Shepp, L. A., & Vardi, Y. (1982). Maximum likelihood reconstruction for emission tomography. IEEE transactions on medical imaging, 1(2), 113-122.

[10] Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., & Webb, R. (2017). Learning from simulated and unsupervised images through adversarial training. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2107-2116).

[11] Zhang, Z., Yang, L., & Zheng, Y. (2018). Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 9242-9251)s.

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

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

[14] Nie, D., Trullo, R., Lian, J., Petitjean, C., Ruan, S., Wang, Q., & Shen, D. (2017, September). Medical image synthesis with context-aware generative adversarial networks. In international conference on medical image computing and computer-assisted intervention (pp. 417-425). Springer, Cham.

[15] Burgos, N., Cardoso, M. J., Guerreiro, F., Veiga, C., Modat, M., McClelland, J., ... & Hutton, B. F. (2015, October). Robust CT synthesis for radiotherapy planning: application to the head and neck region. In international conference on medical image computing and computer-assisted intervention (pp. 476-484). Springer, Cham.

[16] Bahrami, K., Shi, F., Rekik, I., & Shen, D. (2016). Convolutional neural network for reconstruction of 7T-like images from 3T MRI using appearance and anatomical features. In deep learning and data labeling for medical applications (pp. 39-47). Springer, Cham.

[17] Fan, H., Su, H., & Guibas, L. J. (2017). A point set generation network for 3d object reconstruction from a single image. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 605-613).

[18] Choy, C. B., Xu, D., Gwak, J., Chen, K., & Savarese, S. (2016, October). 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In European conference on computer vision (pp. 628-644). Springer, Cham.

[19] Jiang, L., Shi, S., Qi, X., & Jia, J. (2018). Gal: Geometric adversarial loss for single-view 3d-object reconstruction. In Proceedings of the European conference on computer vision (ECCV) (pp. 802-816).

[20] Yang, G., Cui, Y., Belongie, S., & Hariharan, B. (2018). Learning single-view 3d reconstruction with limited pose supervision. In Proceedings of the European conference on computer vision (ECCV) (pp. 86-101).

[21] Xu, Q., Wang, W., Ceylan, D., Mech, R., & Neumann, U. (2019). Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In advances in neural information processing systems (pp. 492-502).

[22] Wu, J., Zhang, C., Zhang, X., Zhang, Z., Freeman, W. T., & Tenenbaum, J. B. (2018). Learning shape priors for single-view 3d completion and reconstruction. In Proceedings of the European conference on computer vision (ECCV) (pp. 646-662).

[23] Fernandez-Labrador, C., Perez-Yus, A., Lopez-Nicolas, G., & Guerrero, J. J. (2018). Layouts from panoramic images with geometry and deep learning. IEEE robotics and automation letters, 3(4), 3153-3160.

[24] Kolev, K., Brox, T., & Cremers, D. (2012). Fast joint estimation of silhouettes and dense 3D geometry from multiple images. IEEE transactions on pattern analysis and machine intelligence, 34(3), 493-505.

[25] Choy, C. B., Xu, D., Gwak, J., Chen, K., & Savarese, S. (2016, October). 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In European conference on computer vision (pp. 628-644). Springer, Cham.

[26] Xie, H., Yao, H., Sun, X., Zhou, S., & Zhang, S. (2019). Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In Proceedings of the IEEE international conference on computer vision (pp. 2690-2698).

[27] Würfl, T., Ghesu, F. C., Christlein, V., & Maier, A. (2016, October). Deep learning computed tomography. In international conference on medical image computing and computer-assisted intervention (pp. 432-440). Springer, Cham.

[28] Hammernik, K., Würfl, T., Pock, T., & Maier, A. (2017). A deep learning architecture for limited-angle computed tomography reconstruction. In bildverarbeitung für die medizin 2017 (pp. 92-97). Springer Vieweg, Berlin, Heidelberg.

[29] Bayat, A., Sekuboyina, A., Paetzold, J. C., Payer, C., Stern, D., Urschler, M., ... & Menze, B. H. (2020). Inferring the 3D standing spine posture from 2D radiographs. arXiv preprint arXiv:2007.06612.

[30] Henzler, P., Rasche, V., Ropinski, T., & Ritschel, T. (2017). Single-image tomography: 3d volumes from 2d x-rays. arXiv preprint arXiv:1710.04867.

[31] Song, W., Liang, Y., Wang, K., & He, L. (2020). Oral-3D: Reconstructing the 3D bone structure of oral cavity from 2D panoramic X-ray. arXiv preprint arXiv:2003.08413.

[32] Kasten, Y., Doktofsky, D., & Kovler, I. (2020). End-to-end convolutional neural network for 3D reconstruction of knee bones from bi-planar X-ray images. arXiv preprint arXiv:2004.00871.

[33] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In international conference on medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

[34] Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., & Paul Smolley, S. (2017). Least squares generative adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 2794-2802).

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

[36] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., ... & Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4681-4690).

[37] Zhu, B., Liu, J. Z., Cauley, S. F., Rosen, B. R., & Rosen, M. S. (2018). Image reconstruction by domain-transform manifold learning. Nature, 555(7697), 487-492.

[38] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

[39] El Khiyari, H., & Wechsler, H. (2016). Face recognition across time lapse using convolutional neural networks. Journal of information security, 7(3), 141-151.

[40] Dumoulin, V., & Visin, F. (2016). A guide to convolution arithmetic for deep learning. arXiv preprint arXiv:1603.07285.

[41] Karpathy, A. (2016). Neural networks part 1: Setting up the architecture. Notes for CS231n convolutional neural networks for visual recognition.

[42] James D. McCaffrey. Convolution image size, filter size, padding and stride. https://jamesmccaffrey.wordpress.com/2018/05/30/convolution-image-size-filter-size-padding-and-stride/ , 2018. (Accessed: 2020-9-29).

[43] Nair, V., & Hinton, G. E. (2010, January). Rectified linear units improve restricted boltzmann machines. In ICML.

[44] Xu, B., Wang, N., Chen, T., & Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853.

[45] Reza, Z. N. (2019). Real-time automated weld quality analysis from ultrasonic B-scan using deep learning.

[46] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

[47] Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2016). Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022.

[48] Armato III, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., ... & Kazerooni, E. A. (2011). The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Medical physics, 38(2), 915-931.

[49] Jaeger, S., Candemir, S., Antani, S., Wáng, Y. X. J., Lu, P. X., & Thoma, G. (2014). Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quantitative imaging in medicine and surgery, 4(6), 475.

[50] Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K. I., ... & Doi, K. (2000). Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. American journal of roentgenology, 174(1), 71-74.

[51] Van Ginneken, B., Stegmann, M. B., & Loog, M. (2006). Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. Medical image analysis, 10(1), 19-40.

[52] Milletari, F., Navab, N., & Ahmadi, S. A. (2016, October). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In 2016 fourth international conference on 3D vision (3DV) (pp. 565-571). IEEE.

[53] Kingma, D. P., & Ba, J. (2014). Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[54] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4), 600-612.

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

[56] Yang, L., & Chakraborty, R. (2020, April). An “Augmentation-Free” rotation invariant classification scheme on point-cloud and its application to neuroimaging. In 2020 IEEE 17th international symposium on biomedical imaging (ISBI) (pp. 713-716). IEEE.

[57] Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7794-7803).

[58] Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., & Chua, T. S. (2017). Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5659-5667).

[59] Jiang, Y., Li, P., Zhang, Y., Pei, Y., Guo, Y., Xu, T., & Yuan, X. (2020, October). 3D volume reconstruction from single lateral X-ray image via cross-modal discrete embedding transition. In international workshop on machine learning in medical imaging (pp. 322-331). Springer, Cham.

[60] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[61] Guan, Q., Huang, Y., Zhong, Z., Zheng, Z., Zheng, L., & Yang, Y. (2018). Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. arXiv preprint arXiv:1801.09927.

[62] Caponetti, L., & Fanelli, A. M. (1990, November). 3D Bone reconstruction from two X-Ray views. In [1990] Proceedings of the twelfth annual international conference of the IEEE Engineering in Medicine and Biology Society (pp. 208-210). IEEE.