UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

RADIOLOGICAL QUALITY ASSURANCE IN DIGITAL CHEST X-RAY IMAGING

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL ELÉCTRICO

TOMÁS ANDRÉS DE LA SOTTA KRAUSE

PROFESOR GUÍA:
JOSÉ MANUEL SAAVEDRA RONDO

MIEMBROS DE LA COMISIÓN:
IVAN SIPIRÁN MENDOZA
PABLO ESTÉVEZ VALENCIA

Este trabajo ha sido financiado por:
RetinaRX

Este trabajo ha sido realizado en cotutela con:
Universidad de los Andes

SANTIAGO DE CHILE
2023

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL ELÉCTRICO
POR: TOMÁS ANDRÉS DE LA SOTTA KRAUSE
FECHA: 2023
PROF. GUÍA: JOSÉ M. SAAVEDRA RONDO

ASEGURAMIENTO DE CALIDAD RADIOLÓGICA EN IMAGENOLOGÍA DIGITAL EN RAYOS-X DE TÓRAX

Casi el 30% de las radiografías tomadas en todo el mundo nunca son vistas por un radiólogo cualificado, lo que supone una enorme disminución en la salud de las personas. Este valor aumenta constantemente con el tamaño de la población, y debe ser contenido.

La inteligencia artificial se ha acercado a este problema de varias maneras, pero aún carece de la capacidad de mejorar la calidad y la eficiencia del trabajo del médico radiólogo.

En el siguiente documento, desarrollamos un sistema basado en IA para determinar los estándares de calidad en imágenes de rayos X de tórax. Para ello, estudiamos los estándares médicos de calidad de imagen, determinando un sistema de garantía de calidad en dos pasos. Dado que la calidad de la imagen médica de rayos X de tórax depende de las estructuras semánticamente visibles presentes en la cavidad torácica, primero estudiamos múltiples variaciones de modelos *U-Net* para la segmentación estructural, seguido de la evaluación de técnicas de *image processing* basadas en la pre-segmentación de órganos para la determinación de la calidad.

Dentro del estudio de modelos de visión por computador, se ha demostrado que las *capas atencionales* mejoran el rendimiento, permitiendo al modelo "elegir dónde ver", aumentando la eficiencia y reduciendo el número de imágenes necesarias en el momento del entrenamiento. Para ello, se estudian varias arquitecturas estado-del-arte de segmentación de rayos X de tórax, basadas en *U-Net* y modificadas con mecanismos *attencionales*, presentando técnicas *convolucionales*, *atencionales* y mixtas para la resolución del problema propuesto. Estos modelos son entrenados y evaluados sobre cuatro datasets públicos independientes, siendo estos, los datasets correspondientes a: Montgomery County, Shenzhen Tuberculosis, JSRT y VinDr-RibCXR.

Para la determinación de la calidad radiológica, se aplican algoritmos de *image processing* a las segmentaciones dadas, los cuales permiten la definición de ciertas métricas de calidad en cada una de las áreas definidas.

Como resultado, se observa una mejora en el rendimiento, precisión y robustez de los modelos de segmentación presentados, junto con una disminución en el número de imágenes necesarias para entrenar estos sistemas. Esta mejora se observa en hasta un 6,4% de incremento en el valor DICE respecto al modelo U-Net entrenado en 222 imágenes de entrenamiento y un 2,9% de incremento para 610 imágenes, mostrando resultados aceptables en la calidad, siendo 48%, 23% y 22% de *accuracy* para las categorías respectivas.

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL ELÉCTRICO
POR: TOMÁS ANDRÉS DE LA SOTTA KRAUSE
FECHA: 2023
PROF. GUÍA: JOSÉ M. SAAVEDRA RONDO

RADIOLOGICAL QUALITY ASSURANCE IN DIGITAL CHEST X-RAY IMAGING

Almost 30% of world-wide taken x-ray images are never seen by a trained radiologist, causing a huge reduction in people's health. This value is steadily incrementing with population size, and must be contained.

Artificial intelligence has approached this problem in several ways, yet lacking the capacity to improve the quality and time-efficiency of radiologist's work.

In the following document, we develop an AI based system to determine the quality standards for chest x-ray imaging. For this purpose, we study the medical standards for image quality, determining a two-step quality assurance system. As medical chest x-ray image quality depends on the semantically visible structures present in the thoracic cavity, we first study multiple *U-Net* model variations for structural segmentation, followed by evaluating organ-segmented based *image processing* techniques for quality determination.

Within the study of computer vision models, it has been shown that *attentional layers* improve their performance, allowing the model to "choose where to see", increasing efficiency and reducing the number of images needed at the time of training. For this purpose, several *U-Net* based *state-of-the-art* x-ray architectures are studied and modified with *attentional* mechanisms, presenting *convolutional*, *attentional* and mixed techniques for the resolution of the proposed problem. These models are trained and evaluated on four independent public datasets, being these, the datasets corresponding to: Montgomery County, Shenzhen Tuberculosis, JSRT and VinDr-RibCXR.

For the determination of radiological quality, *image processing* algorithms are applied to the given segmentations, which allow the definition of certain quality metrics in each of the defined areas.

As a result, an improvement in the performance, accuracy and robustness of the presented segmentation models is observed, together with a decrease in the number of images required to train these systems. This improvement is observed in up to a 6.4% increase in the DICE value over the U-Net model in 222 training images and a 2.9% increase for 610 images, showing acceptable results for quality assurance of 48%, 23% and 22% *accuracy* for each of the quality assurance categories.

Una vez que la tormenta termine, no recordarás cómo lo lograste, ni cómo sobreviviste. Ni siquiera estarás seguro de si la tormenta ha terminado realmente. Pero una cosa sí es segura. Cuando salgas de esa tormenta, no serás la misma persona que entró en ella.

De eso trata la tormenta

Haruki Murakami

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1. Introduction

Medical imaging is an area of medicine that focuses on image-based diagnosis. This area has allowed physician to improve diagnosis and treatments through non-invasive procedures. Some popular medical imaging modalities are X-ray projection radiography, X-ray computed tomography (CT), magnetic resonance imaging (MRI), nuclear imaging like SPECT and PET, fluoroscopy, among others [1].

X-ray imaging is considered the oldest medical imaging modality that appeared after the Roetgen's discovery at the end of the 19th century [2]. The X-ray procedure is simple and provides images directly, without any reconstruction or costly methods.

In this vein, chest X-ray imaging is the most common image-based means for probing pulmonary disorders. For instance, tuberculosis, tumors, pneunomia, fibrosis and pulmonary nodules are prime examples of disorders that can be detected by a chest X-ray procedure. According to Broader [3], chest X-ray is one of the most cost-effective and non-intrusive imaging examination. Through a chest X-ray physicians can inspect organs like heart, lungs, ribs, bones or even blood vessels.

Although X-ray is the lowest-cost imaging, its capabilities are far from low when studied by a trained radiologist, giving the possibility of detecting complex diseases with a fast and almost cost-free exam. This allows us to enhance patient's experience, reducing time, cost and, over all things, complex machinery, being able to get good results in almost every medical center conditions.

Regrettably, this type of exam is considerably overused, presenting over 5 million chest x-ray images taken in Chile the year 2017, being approximately 27% of the country's population. Considering that in 2017 there were around 1,100 expert radiologists, the problem arises. In world-wide considerations, around 30% of x-rays are never seen by a trained radiologist, which causes a decrease in people's health due to thousands undetected conditions.

The principle of X-ray images is based on the level of absorption of the organs when a X-ray beam passes through them. Dense organs present more absorption levels than lesser dense ones. Thus, high-density structures, like bones, appear white in the X-ray image, while

soft organs, like lungs, appear in a gray color scheme [4].

Although an X-ray examination is a quick procedure, it requires an expert to interpret the image. The inter-patient variability and the fuzzy delimitation of organs makes the image interpretation a difficult task. In addition, the workload of experts can increase the chance of error during this process.

Considering that the images are taken by a manually-configured machine on unprofessional x-ray model patients, the quality of the outputted image can be considerably affected. Low quality images can reduce the inner-chest visibility as well as compressing the organs, leading to wrong diagnosis or hidden conditions that remain unseen.

A critical stage in image analysis and quality assurance is image segmentation. The goal of segmentation is to delimit different meaningful parts of an image at a pixel-size level. In the case of medical images a segmentation method should delimit organs, bone tissues and vascular structures [2]. As a result of structural x-ray segmentation, physicians are able to provide more accurate diagnosis or efficient radiotherapy plans. In fact, lung segmentation in chest x-ray imaging is critical for cancer diagnosis [5].

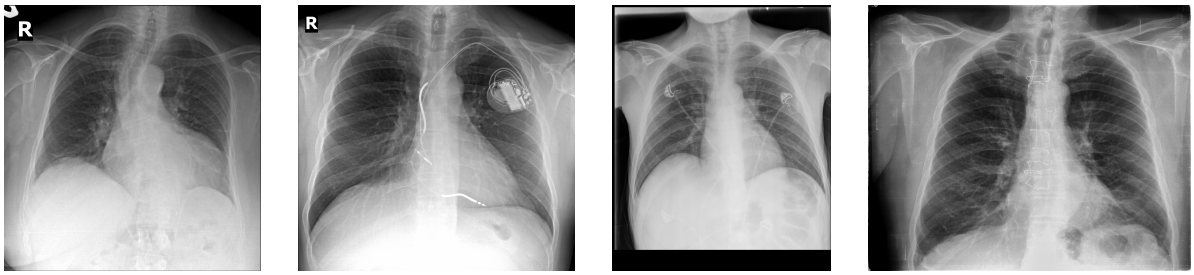
Computational advances have made possible to generate new solutions for the adversities mentioned above, especially in matters related to artificial intelligence and image applications; which currently work as an assistant for the specialist both in the scientific area (investigating causes and analyzing results) and the medical area (when supporting diagnoses and treatments). In this sense, the significant advances in deep learning and computer vision, have allowed to find more precise and valuable characteristics, and patterns, that provide clues or conclusions about an image's context.

In the extent of this document, we will study different U-Net variants for determining clavicles, ribs, heart and lung segmentation masks, with the objective to define a system that provides quality assurance advice for digital chest x-ray imaging.

1.2. The Quality Assurance Problem

Medical image quality assurance is not as simple as traditional image quality assessments. This complexity is due to the importance of image semantics in the medical field, leaving the remaining factors; such as contrast, saturation, focus, color, etc.; in a *not less valuable* second plane.

As we work specifically with ambulatory chest x-rays, it is really important to know the difference and characteristics of this specific imaging segment.
First of all, ambulatory imaging is based on the assumption that the patient is able to be positioned in what is called *bipedestation*, or *standing* in a more common and, almost, oversimplified manner. This assumption is crucial for this document, because most of the organs inside our body are made from soft tissue, which deforms differently, depending on the patient's position. This natural deformation could, for example, show signs of cardiomegaly; the enlargement of the heart; in posteroanterior chest x-ray imaging on lying patients.



(a) Severe case of scoliosis

(b) Pacemaker

(c) External chest electrodes

(d) Sternal wiring

Figure 1.1: Medical artifacts in chest x-rays

The second ambulatory assumption consists on non-artifacts carrying patients. These artifacts consist, principally, on sternal wires, pacemakers, electrodes and other medical devices.

Finally, the third ambulatory assumption consists on a reduced condition list, as there are many conditions that will never be present in ambulatory cases due to their complexity or the urgency. For this point is important to highlight that emergency room cases are not considered ambulatory in medical terms, so the system will never encounter some severe situations.

1.2.1. Quality Assurance Factors

In medical imaging, there are three main factors for determining quality: the radiation penetration, the patient rotation and the lung insuflation.

1.2.1.1. Radiation Penetration

One of the most tackled problems in AI for medical image quality assurance is the radiation penetration or image hardness. This problem is the simplest of the three presented in this section, because is highly related to the classical contrast-ratio and image-histogram quality problem.

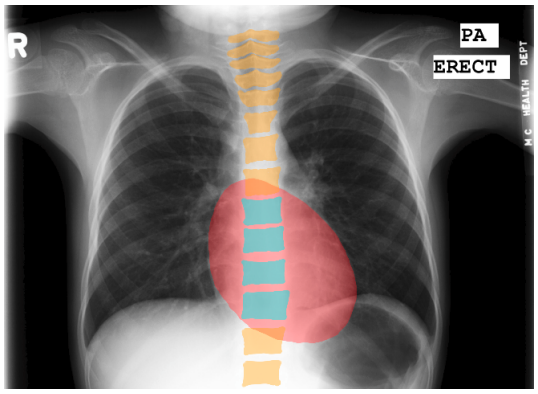


Figure 1.2: Correctly Penetrated X-Ray Visualization

For the extent of this document, we will look to the radiation penetration problem as the number of thoracic vertebrae visible behind the heart's shade, as detailed by the radiologists team associated with RetinaRX. Three to four vertebrae are considered ideal for classical chest x-ray images, as shown in fig. 1.2.

1.2.1.2. Patient Rotation

The patient rotation is one of the key aspects in radiological quality assurance. It ensures that the image projection is correctly positioned in the frontal plane, reducing at it's maximum, the visible organs superposition. As an ambulatory chest x-ray patient has to be standing, we study only the longitudinal (vertical) axis rotations. The other two rotational axes imply the patient inclined or laying down, both of which cases are excluded in the base assumptions of this document.

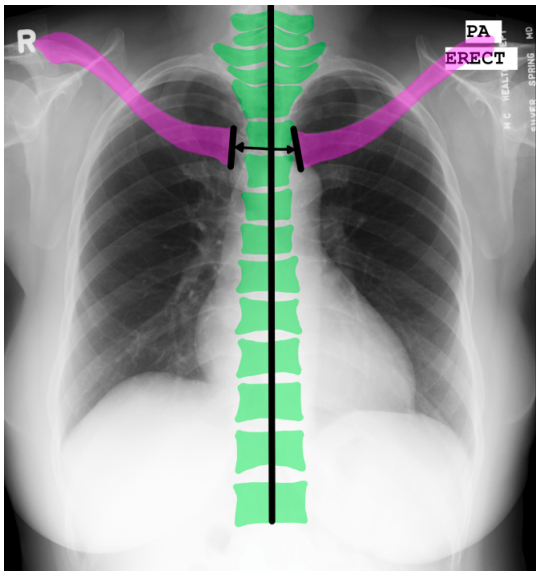


Figure 1.3: Slightly Rotated Image Visualization

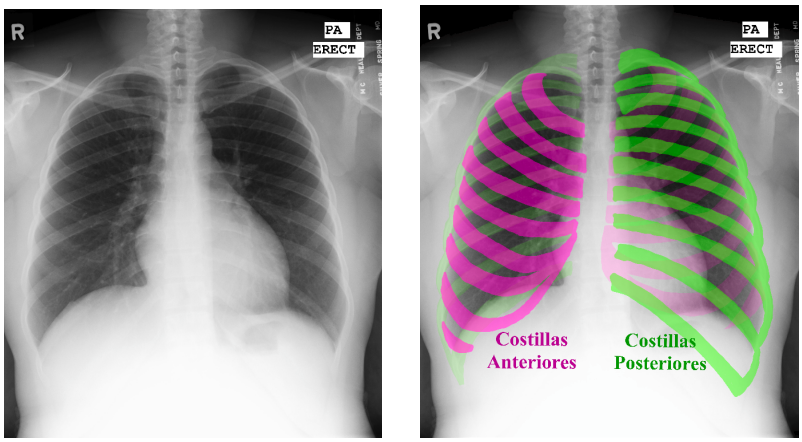
For determining the rotation of an image, it is key the study of the clavicles themselves. As there are various patients presenting scoliosis or curving, the rotation is studied over a single, fixed and invariant point: the clavicle distance from the medial plane. This metric measures the distance relationship between each clavicle's head and the medial plane, described as the division of both gaps. A correctly rotated image should present this metric in the vicinity of 1. Larger or smaller numbers outside of a tolerated threshold are to be considered rotated and should be retaken.
The medial plane can be determined by the vertebral column, the sternum, the ribs or the whole thoracic cage. In fig. 1.3, the medial plane is determined by the column itself. Cases of scoliosis are complex situations for this medial plane determination way.

1.2.1.3. Lung Insuflation

The amount of air holded by the patient during the image acquisition is called *lung insuflation* and determines the amount of compression in the thoracic cage. If the lungs are overinsuflated, the organs will be compressed against each other. If the patient underinsu-

flates his lungs, the thoracic cage will be shrinken down, resulting in the same problem as overinsuflation: compression.

For determining the lung insuflation over an image, the proceeding is quite simple: counting visible ribs. Human beings have around 12 ribs in each thorax side, 10 *true* ribs that converge in the sternum, and 2 *floating ribs* that are simply linked to the vertebral spine. This set of ribs is medically subdivided in two sections: anterior ribs and posterior ribs. Anterior ribs count a total number of 10 different bone structures, all linked to sternum. This set of ribs correspond to the frontal section of the thoracic cage, while divided by the coronal plane. The posterior ribs are exactly the opposite, being the section of ribs to the back of the same plane. As posterior ribs consider the two floating ribs, the number of total posterior ribs adds up to 12, two more than the anterior case.



(a) Correctly Lung-Insuflated Image

(b) Anterior and Posterior Ribs Manually Segmented

Figure 1.4: Lung Insuflation Visual Representation

An ideal mark for insuflation quality assurance consists on counting 6 visible anterior ribs or 10 posterior ribs with the same characteristics, as shown in figure 1.4.

1.3. The Architectural Background

In the extent of this section, we will develop the basic knowledge around the theory used for developing this document. It is important to note that this section is not an introduction to deep learning, but is a simple, reduced and explained background for understanding the basic details of the used architectures.

1.3.1. A Brief Explanation of Deep Learning

It is usually misunderstood the way machine learning works, thinking it similes a human brain, having a curiously mystic and magical idea over them. Lamentably, this is not the way AI systems work, so lets break the idea on how they do work.

1.3.1.1. Linear Layer

One of the first models called *neural* are basically what is known today as *Multi-Layer Perceptron* or MLP. These models are basically a set of *learnable* parameters, or weights, that ponderate the input values, independently, to create a determined-size output. For example, given an input vector of size 3 and a desired 2-size output vector, we define the MLP layer as the set of weights A given as shown in equation 1.1.

$$A = \{a_1^1, a_1^2, \cdots, a_2^3\} \tag{1.1}$$

This weights add and multiply each point of the input vector for getting each output. Note that we defined the names of the weights with an subscript, defining the input parameter to which it multiplies, and a superscript revealing the output element where added. This application is shown in equation 1.2, where the input vector is a 3-size vector defined as $[x_1, x_2, x_3]$.

$$output = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} x_1 \cdot a_1^1 + x_2 \cdot a_2^1 + x_3 \cdot a_3^1 \\ x_1 \cdot a_1^2 + x_2 \cdot a_2^2 + x_3 \cdot a_3^2 \end{bmatrix} \tag{1.2}$$

To simplify the notation, we redefine A as a matrix, by arranging the values as shown in equation 1.3

$$A = \begin{cases} \begin{bmatrix} a_1^1 & a_1^2 \\ a_2^1 & a_2^2 \\ a_3^1 & a_3^2 \end{bmatrix} \end{cases} \tag{1.3}$$

Using this notation, we can re-write the output vector V in a matrix equation as shown in 1.4, where V is the output vector, X the input vector and A the redefined weight matrix.

$$V = X \cdot A \tag{1.4}$$

This linear operation allows to have *learnable* parameters, which correspond to the matrix A. The problem behind this operation is that most of the tasks are not linear, so they can't be modelled by this solely equation. For resolving this issue, a non-linearity function f is added, so it transforms a linear space to a non-linear space. We also add a bias vector B, that corresponds to another set of weights that adds to the output, giving more information to the layer. The final non-linear layer, called *Fully Connected* (FC), is shown in equation 1.5. A stack of consecutive FC layers is usually called Multi-Layer Perceptron (MLP[1]).

$$V = f(X \cdot A + B) \tag{1.5}$$

1.3.1.2. Learning? No, Optimization

Having the mathematical structure of a simple network, the question of how it learns it's parameters is implicitly posed. As said in prior sections, the model does not learn magically,

[1] MLP refers to a network type built only by a set of FC layers, with their corresponding non-linearities, being a specific case of Feed-Forward Networks

instead it solves the *Descending gradient Problem*, a method of mathematical optimization for unknown vector spaces.
The idea behind the problem is to determine the best set of weights possible for resulting in the lowest error attainable. Usually, a lot of mathematical problems can be modelled as a function or an equation, but, regrettably, some of them are not possible to approach in this manner.
For this method we will use a simple analogy. Imagine you want to get to the lake, near the local hill in where you are vacationing, with your eyes closed. As you are incapable of seeing, you decide to walk down the hill, hoping the lake will be in the bottom-most part of it. So the first thing you do is walk a single step towards any direction to see if you are going upwards or downwards. If you are going upwards, you turn around and try another step in that new direction. If you are going downwards, you take another step in that direction and evaluate again. This way you know that, in each step, you are assuring you continue to go down the hill, hoping you get to your desired place.
As seen in the analogy, the gradient descent algorithm does exactly the same: it determines the most-decreasing direction via the gradient operator, and takes a step in that direction. In this case, the step size is determined by a parameter called *learning rate*. This way, gradient descent can be defined as shown in equation 1.6, where the function F represents the following layers parameters.

$$X_{n+1} = X_n - l_r \cdot \nabla F(X_n) \tag{1.6}$$

To determine the gradient, the *Loss Function* is a defined metric which determines how far the result is out from the desired output. As the model is composed as a stack of layers, the gradient in each of them depends on the loss function's derivative, the non-linearities, and latter layer's weights.

1.3.2. Convolutional Neural Networks

Linear layers require at least $n \times m$ parameters, where n is the size of the input vector and m, the size of the output, considering no use of bias. If we consider a color image of 512×512 pixels, an output of size 1024 and an standard *float32* codification, each layer will have around 805M parameters, weighting around 3GB of memory without compression. A single layer of 3GB is extremely heavy, specially considering that models, today, often present more than 50 layers in depth. For this reason, in image modelling, one of the most used architectures is the CNN or *Convolutional Neural Network*, which is constructed over low-weight layers, called *Convolutional Layers*.
Convolutional layers are based over a moving window concept, allowing a single, and small, matrix to move around the whole image, weighting every single pixel without the need of a high parameter count.

1.3.2.1. The Convolution Layer

Based over the concept of the mathematical convolution operator, the convolutional layer consists on a moving, odd size, window, weighting and adding the values inside the space it covers at each position. This movement is determined by a *stride* value, that represents how many pixels the window moves at each step, and a *padding* value, that adds a border around the image. The window presented in the convolution corresponds to a matrix that

is point-to-point multiplied and then added together, to reveal the value positioned at the center of the window in the new image. To avoid the reduction of size, it is usually used a *padding* value determined by the floor division by two of the kernel matrix size, for a stride of one.

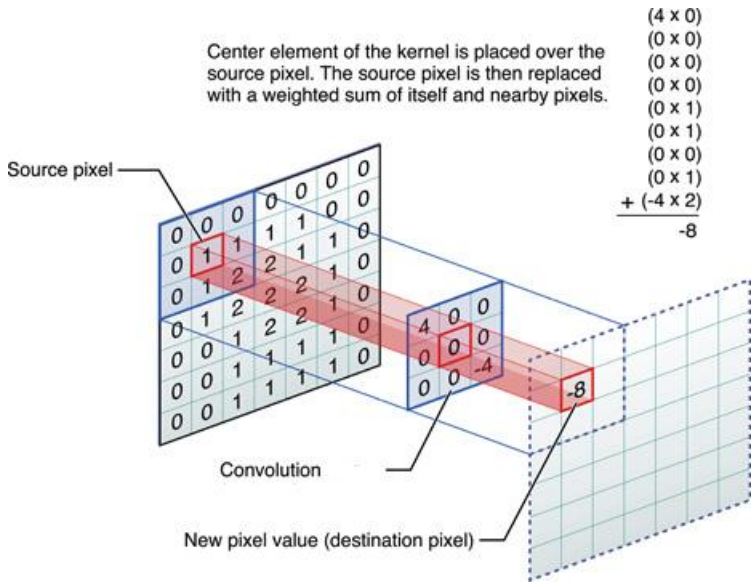


Figure 1.5: Convolutional Operator [6]

1.3.2.2. The Residual Block

A common problem in deep neural networks is that the gradient itself, while propagated through the networks, will lose it's value and become negligible. In deeper models, the step-size variation can be, practically, null, reducing the capacity of the model to learn on deeper layers. This problem is called *The vanishing gradient problem*.
To reduce this problem, the authors of ResNet; a model that we will tackle later on in this document; proposed to define a bypass every determined set of blocks, as shown in fig. 1.6. This method is implemented by outputting the addition of the input and the output of the layer set.
Mathematically, as the gradient propagates backwards on the network, this *bypasses* will act as an identity layer, from which the gradient will present small variations given by the propagation of the layer set inside the block. This method increases the gradient's scope in high-count layer models, allowing deeper structures to learn better, faster, and with less data needs.

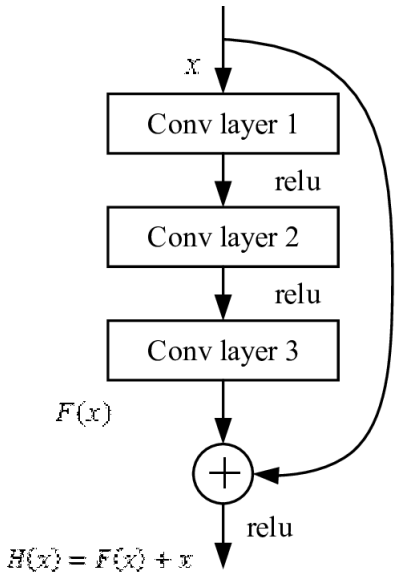


Figure 1.6: Residual Block [7]

1.3.3. Attentional Models

In the development of element-sequence capable models, natural language processing (NLP) has been one of the most remarkable research areas. During these years, models based on recurrent neural networks (RNN)[1] were the basal structure on this area. These models, as the data was sequentially added to the previous information, lose, at each step, information about the preceding elements in the sequence. With the objective of solving this problem, an idea based over a quasi-linear layer that considered every previous output, took a valuable place. The idea proposed allowed the model to to pay attention, choosing the valorization of each step, somehow solving this problem; due to this, the model's structure was called "attention". In 2017, a group of investigators leaded by Ashish Vaswani, showed that these layers could dispense of the RNN, generating a new, better model by themselves, the Transformer.
As for today, attention is a critical component of visual systems. This is close to the *Gestalt Principles* [8] that make some objects be perceived as a whole, facilitating the visual perception. This phenomenon is also called *Perceptual Grouping* and suggests that different receptive fields of an scene influence the others. In computer vision, perceptual grouping can be implemented by attention mechanisms like the popular Transformer modules presented by Vaswani [9].

1.3.3.1. The Attention Block

Fig. 1.7 illustrates the mechanism of self-attention for visual recognition tasks using the Transformer's strategy. As described by Vaswani et al. [9], a transformer module receives the input as a sequence of embeddings. For image tasks, the embeddings can be obtained from a feature map [2] produced by a convolutional neural network [10].

[1] RNN models are a concept of building networks with a fixed set of trainable parameters, over which, sequentially ordered input segments are recurrently complemented to the data over a single layer.

[2] A *feature map* refers to a set of tensors, usually matrices or vectors, that represent some information in a different vector space. In this case, an image feature map is a set of lower pixel-count matrices obtained by convolutional transforms applied to the original image.

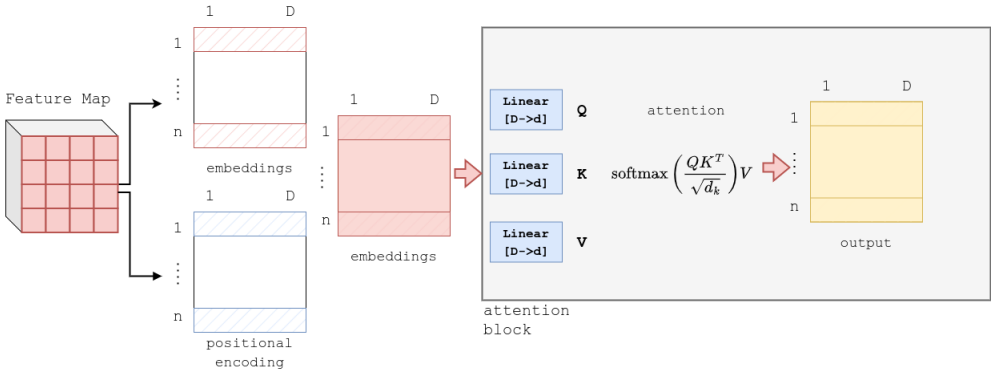


Figure 1.7: Attention mechanism

In addition, we can compute positional encoding that are summed with the visual embeddings. The result is then used as input to the attention block. The attention block requires three representations computed by linear transformations from the input embeddings. These representations are Q, K and V, where the letters come from "Query", "Key" and "Value", respectively. When these three representation comes from the same source, the attention is called "self-attention", but when Q come from a source different from that from where K and V are produced, the attention is called "cross-attention". In any case, after obtaining Q, K and V, the attention is computed by the Equation 1.7, where d_k is a scalar factor given by the dimension of the input tensor.

$$attention(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{1.7}$$

Multi-Head Attention models are mostly used, being these, comprised by a linearly-independent stack of attention modules output-concatenated [9].

1.3.4. The Segmentation Problem

One of the most known computer vision tasks is segmentation. As classification searches to determine a classes for each image, segmentation tries to define the object's form and position. Fig. 1.8 shows that, given an input image, the segmentation model returns the set of pixels associated to the detected object.

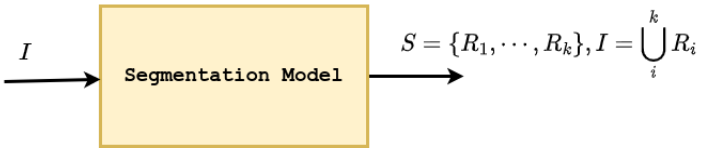


Figure 1.8: Scheme of a segmentation task, where an input image is decomposed into k parts.

Unlike object detection, segmentation aboards a pixel-wise detection, whilst object detection retrieves the object's bounding box. This problem is considered one of the most complex in the computer vision area. The differences between segmentation and object detection are shown in fig. 1.9.

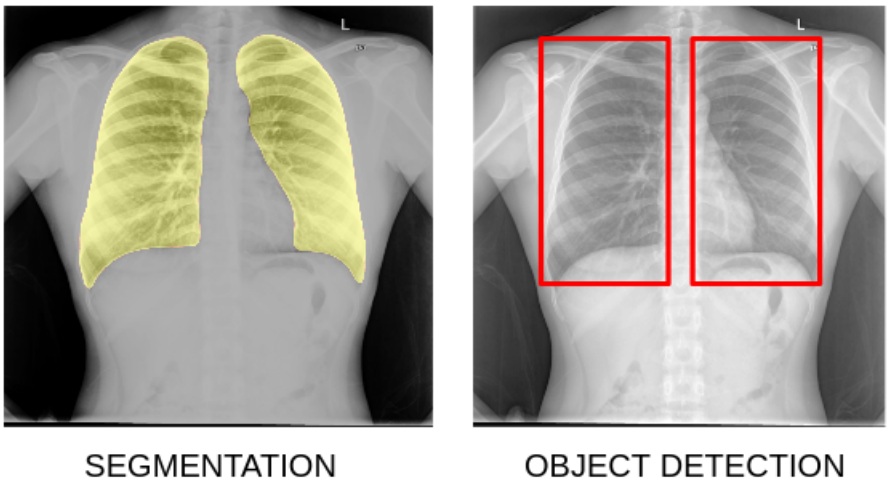


Figure 1.9: (a) Segmentation of Lungs, (b) Object Detection of Lungs

1.4. Base Models

1.4.1. U-Net

U-Net [11] is the most popular deep-learning segmentation model for medical images. It consists of a three-part convolutional neural architecture; the first block is called *the encoder*, which is in charge of computing relevant features from the input image to facilitate the segmentation. Like most neural models, the encoder is a multi-resolution architecture, starting from a high-resolution representation with poor semantic features to lower resolutions with more discriminant features. The encoder is also known as the backbone in a diversity of tasks.

Different from a traditional classification task, where features are aggregated into coarse levels, in the segmentation problem, we need to classify at the finest levels, almost at a pixel level. Therefore, the model needs to take the features produced by the encoder to generate segmentation masks from coarse to the finest levels. To this end, U-Net uses a second block called *the decoder*, that combines information from deeper decoder's layers with higher-resolution encoder levels, allowing the model to generate high-resolution segmentation masks. Figure 1.10 shows a scheme of the U-Net architecture.

Finally, the third block is implicitly positioned, but will take high importance for this document. It is called *skip layers*. The skip layers transport the information gathered by each encoder layer, to the corresponding decoder layer. This is simply done by an identity layer, but will be thoroughly modified and studied.

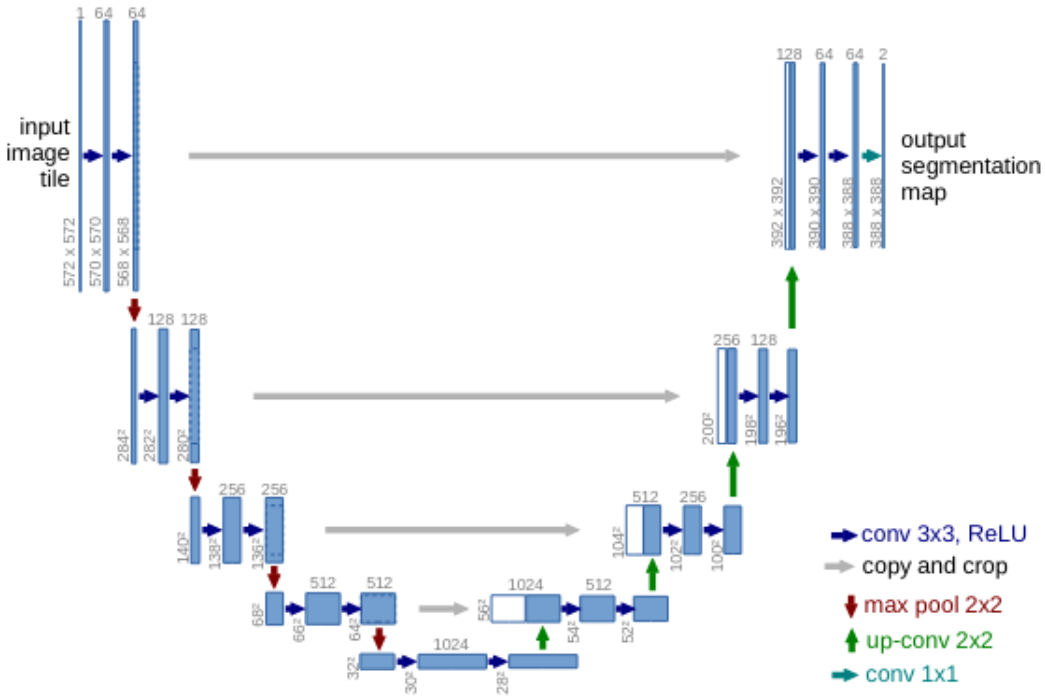


Figure 1.10: UNet Architecture [11]

1.4.2. ResNet

As detailed in section 1.3.2.2, the vanishing gradient problem was, and still is, one of the biggest barriers in what is called *Deep Learning*. For this same reason, ResNet model, presented in 2015 proposed the residual block, shown in figure 1.6, which allowed for the first time to get networks over 50 layers in depth.
ResNet presented multiple models, but for this document, three of them are studied: ResNet-18, ResNet-34 and ResNet-50.

1.4.2.1. ResNet Architectures

ResNet-18 model is an 18-layer depth network with residual blocks bases over two 3x3 consecutive convolutional layers with ReLU and a residual connection. ResNet-34 presents the same structure, but with 34 layers in depth. Both of these models output a standard 1024-size vector for the ImageNet dataset.
ResNet-50 presents a different -and more interesting- approach. This model has 50 layers in depth, with a modified residual block, known as "*bottleneck* residual block", comprised by a dimensional reduction with a 1x1 convolution layer, a characteristics extraction 3x3 convolutional layer, and a dimension-restoring 1x1 convolution layer. All the ResNet model architectures are shown in figure 1.11.

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
		3×3 max pool, stride 2				
conv2_x	56×56	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Figure 1.11: ResNet Architectures [12]

Deeper models as ResNet-101 and ResNet-152 are 101-layer and 152-layer networks based over the bottleneck residual blocks presented in ResNet-50.

1.4.3. Swin Transformer

As detailed in section 1.3.3, transformers are a critical component for computer vision, but they lack of high-resolution capabilities due to their size and memory use.
Multiple models have presented *nouvel* image-transformers architectures, but two of them stand out for computer vision engineers: the *Vision Transformer* [10], ViT; and the Swin Transformer, [13].
ViT presents a patch-embedding structure to feed the information to the original NLP transformer architecture, but this model's complexity escalates extremely fast over a variation on the input image size. For segmentation tasks, this model has not been so fruitful due to the 16x16 pixel partitions that, sometimes, are too big for pixel-wise detailed information needs.
Distinctively to ViT, Swin transformer's block architecture is based on patch-partitioning, residual connections, basic MLPs and a new attention module called *Shifted Window Multi-Head Self Attention* or SW-MSA[3]. This module can use smaller patches than ViT and, as its attention is limited to a small set of patches, the architecture is capable to extract pixel-level characteristics without model-size escalation over image-size.

1.4.3.1. SW-MSA Module

The Self-Attention module presented by the Swin Transformer consists on the application of a *multi-head self-attention* block over small subsets of patches, guided by a moving window strided by the window size, as shown in figure 1.12. This shifting allows the model to capture patch-edge information over standard partitioning.

[3] When shifting is deactivated, the module is usually refered as W-MSA.

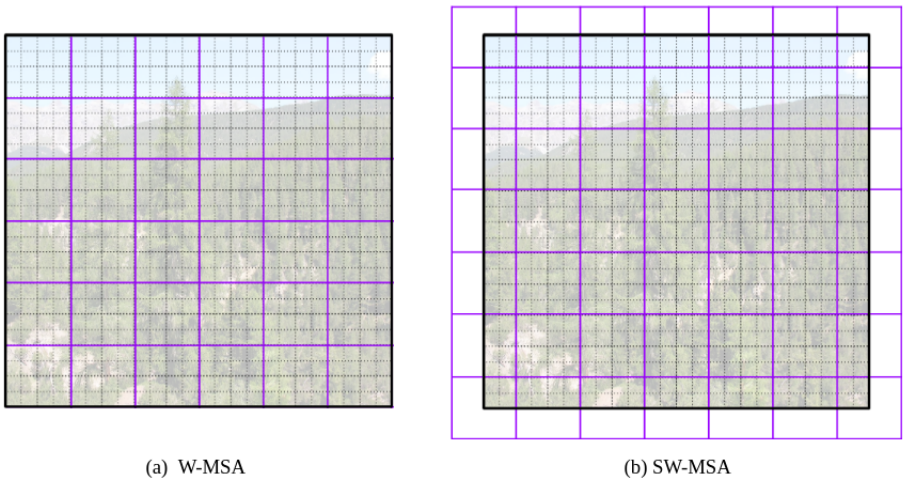


Figure 1.12: Illustration of Shifted Windowing. In (a), a regular W-MSA window partitioning is used, while in (b), the window partition is shifted as used in SW-MSA. Note that the window is represented as a purple square, computing self-attention over a set of patches.

The small patch subsets allow the model to be relatively small compared to a full-set sized transformer block.

1.4.3.2. Swin Block Architecture

The Swin Stage is conformed by patch partitioning and multiple-of-two set of Swin blocks, each one comprised by a SW-MSA module, an MLP module, and two linear layers.
The Swin Stage architecture has a three-step structure. The first step consists on a 4x4 patch partitioning applied to the input. Because of the small window used, this allows the model to observe pixel-size details. Considering a 3-channel RGB image, each patch will represent a tensor of size $4 \times 4 \times 3 = 48$ pixels.
The second and third steps consist on a repeating pattern. Each one of these has two consecutive blocks, where the first one is comprised by two residual connections, one over a linear embedding module and a W-MSA layer, and the other one by a linear embedding and a simple MLP. The linear embedding transforms each patch to a C-dimensional space, with C determined as a model's hyperparameter. The W-MSA layer is applied over the embedded patches to extract the latent characteristics over the first part of the block. The GeLU-based 2-layer MLP step reaches the output residual connection of each transformer block. The Swin Transformer's architecture is shown in figure 1.13.a, where in figure 1.13.b the two-block structure is mostly detailed.

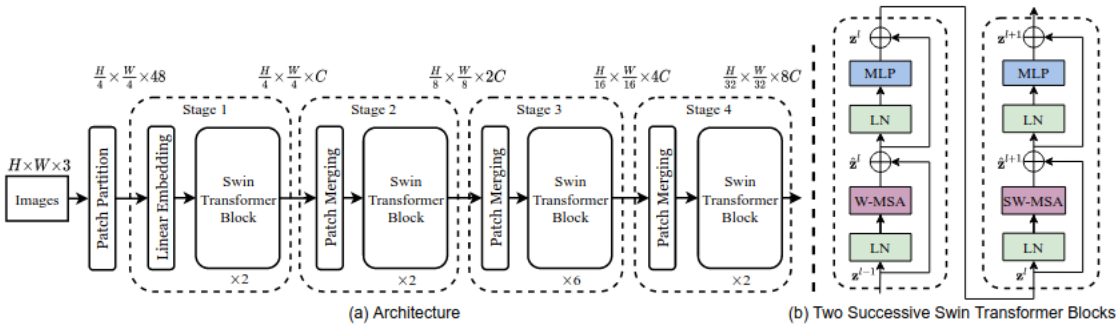


Figure 1.13: Swin Transformer Architecture [13]. (a) Swin Transformer Architecture (Swin-T). (b) Swin Transformer two-block sequential architecture.

As mentioned before each stage present a multiple-of-two set of blocks, so the second sub-block can replace the simple W-MSA with an SW-MSA. This particular change allows the model to detect objects and characteristics that were previously split by the patch partitioning. Each of this multiple-of-two set of blocks are divided into the architecture presented in 1.13.b.

1.4.4. BoTNet: Bottleneck Transformers

Another transformer-based approach taken in computer vision is implied in BoTNet [14], a pseudo-attentional model based on ResNet and the attentional operator to develop a backbone-style semi-convolutional network. This model utilizes the structure provided by ResNet-50, replacing the central 3x3 convolution by a new pseudo-attentional module, inaccurately called *BoTNet Multi-Head Self Attention* or MHSA. For the development of this document we will use the term *Attention* indistinguishably for pseudo-attentional and real-attention based models, this does will not affect the comprehension of the models and, if needed, will be pointed out and specified.

1.4.4.1. BoTNet Multi-Head Self Attention Module

The module presented in BoTNet is heavily used in the development of this work. It is based over the classical attention mechanism showed in equation 1.7, replacing each input's linear layer with simple 1x1-sized convolutional ones, as shown in figure 1.14.

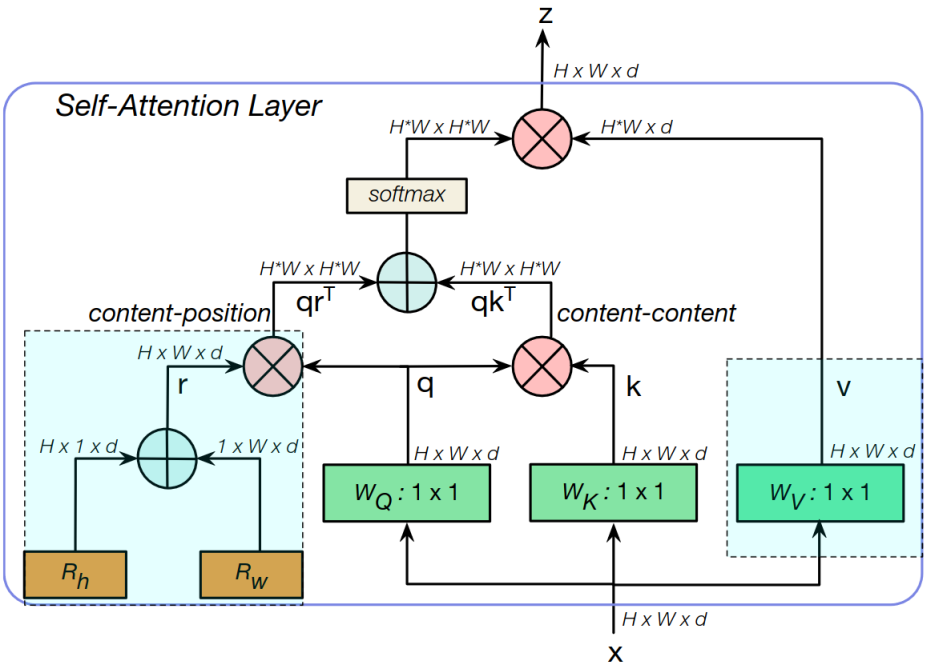


Figure 1.14: BoTNet Multi-Head Self Attention module architecture [14].

A positional encoding layer is added to the query Q value.

1.5. Loss Functions and Metrics

As detailed in section 1.3.1.2, loss functions determine the network step-variation needed for learning, being a hugely important and critical part of the training procedure. In this brief section, the used loss function and the main evaluation metric will be explained.

1.5.1. Dice-Sørensen Score

Given a hand-made segmentation, there are various methods to evaluate the quality of image segmentation. In general, the idea is to measure the difference between the automatic segmentation S against the manually segmented image G by computing some evaluation metric. These metrics can be based on spatial overlap measures (e.g., Dice coefficient [15]) and on distance measures (e.g., Hausdorff distance [16]).

Our evaluation metric is similar to the one used in previous works related to organ segmentation, that is, the Dice coefficient, to compare our results to the ones of the state-of-the-art methods.

The Dice coefficient value is calculated as shown in eq. 1.8.

$$D = \frac{2|S \cap G|}{|S| + |G|} \tag{1.8}$$

where S represents an automatic segmentation method, and G represents the hand-segmented masks (gold-standard). Therefore, the Dice coefficient values vary in the range $[0, 1]$, where 0 indicates no spatial overlap between S and G, while 1 indicates complete overlap.

1.5.2. Cross-Entropy Loss

Given a manually-classified input tensor, the determination of the loss, or sometimes mis-called *distance*, is key for deriving the gradient value for training. Multiple loss functions have been proposed, but one of the most used is the Cross-Entropy loss function.
Derived from the maximum-likelihood estimator, the log-like cross-entropy loss has shown to best perform most of the traditional loss functions in classification tasks. This function is structured over the logarithmic probability for each class, weighted by the true label of itself, as shown in equation 1.9.

$$\mathcal{L}_{CE}{}^{4} = -\sum_{i=1}^{n} t_i \log(p_i) \qquad \text{for n classes} \tag{1.9}$$

[4] Cross-Entropy Loss, where t_i corresponds to the true label and p_i represents the Softmax probability for the i^{th} class. Note log equivalent to base 2 logarithm $\log_2$ for this representation.

Chapter 2

Related Work

2.1. Image Segmentation

For image segmentation, the first computer-based segmentation method for medical images were based on low-level features like pixel intensities or colors and mid-level features like local patterns, but none to them analyzes the semantic content of the images [17]. More close to a semantic approach are the ATLAS-based models that take into account anatomy information [18].

However, with the explosion of deep learning in different computer vision tasks, models have achieved outstanding performance. This is the case of image segmentation, where models leverage labeled images (segmentation masks) to make a model learn a mathematical approximation function that discriminates pixels among a fixed set of classes.

In the medical context, U-Net [11] is the most popular deep-learning architecture proposed for segmenting medical images. It generalizes a previous segmentation model named FCN (Fully convolutional network) [19] that combines features from high resolution layers with those from deeper ones.

In recent years, attention mechanisms have attracted the interest of the community, specially after the positive impact in natural language processing [9]. In visual perception, there is a phenomenon called "perceptual grouping" where various elements in a complex display are perceived as going together in one [20]. The interrelation of different visual components (a.k.a. visual structure) may be learned from experience, particularly during the first months of our lives. Regarding the relevance of perceptual grouping for a visual system, it also can bring improvements in medical image segmentation.

Even though, attention is a popular mechanism in natural language processing, the effectiveness in medical image segmentation has not been deeply studied. Therefore in this work we evaluate the impact of attention mechanisms for multiple chest organ segmentation. We incorporate attention modules in a U-Net architecture under different settings.

2.2. Chest X-Ray Structure Segmentation

U-Net has been widely studied in the context of medical images. For instance, [21] proposed an improved U-Net for lung segmentation. They improved U-Net by using the pre-training Efficientnet-b4 as the encoder and residual blocks and LeakyReLU activations in the decoder.

In [22], the medical-image implementation of ResNet-UNet model [23] is another U-Net improvement, replacing the model encoder with different ResNet variations.

As seen in [24], U-Net++ variation was evaluated in the Montgomery-Shenzhen compound dataset for lung segmentation, providing state-of-the-art results in the defined task.

To improve diversification in the attention mechanism, Vaswani et al. [9] proposed a multi-head attention module, that applies multiple single attention over the input sequence of embedding, independently. All the attentional outputs are then combined by concatenation, and passed through a linear transformation. Figure 2.1 illustrates an scheme of a three-head attention module, as proposed by [9].

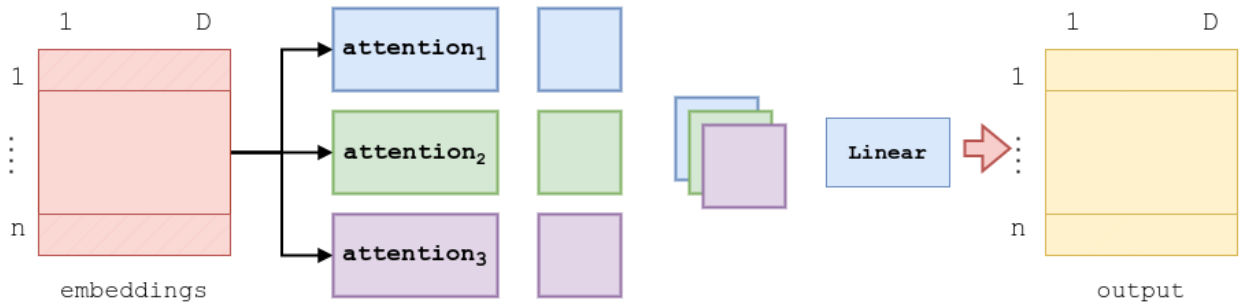


Figure 2.1: A Three-Head Attention Module [9]

In terms of attention regions, [25] proposed a three-input channel-wise attention mechanisms for lung segmentation, combining features from both the encoder and the decoder. Fig. 2.2 depicts the attention module proposed in [25]. However, this proposal does not leverage the interrelation between different receptive field as modern attention Transformers do.

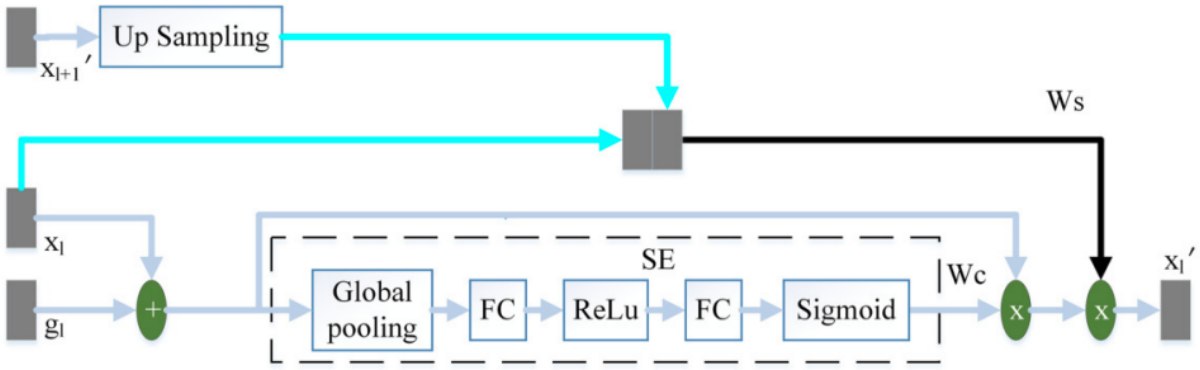


Figure 2.2: A three-input attention proposed in [25] for lung segmentation.

In [25], [26] presents another pseudo-attentional U-Net based model is presented, evaluating the use of attentional feature-extraction and feature-merging in encoder and decoder layers, as shown in figure 2.3.

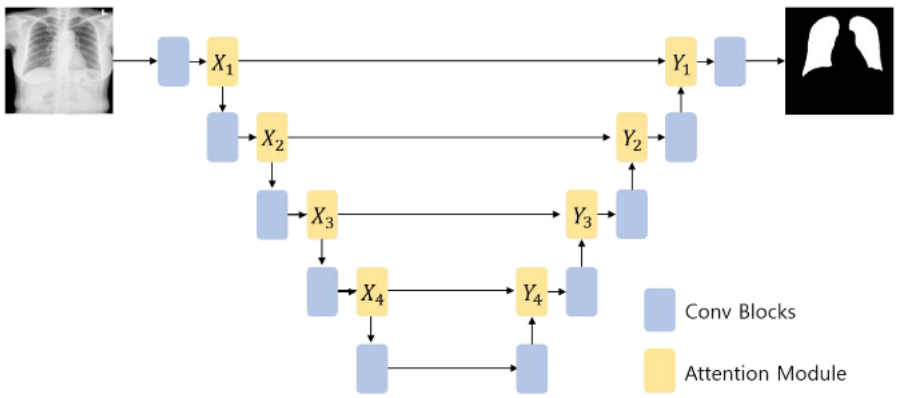


Figure 2.3: Full Application of X-Y Attention Modules [26].

Another lung segmentation approach evaluated the use of variational data inputation over the U-Net base model [27], showing an improvement over the base model, but not achieving state-of-the-art results.

Finally, some non U-Net-like approaches have been evaluated, such as [28] that uses SVMs and Active Shape Models, *ASM*, to segment the lung in thoracic PA and AP images.

For ribs, heart and clavicles, the segmentation models are vaguely approached. In [29] the MDU-Net is proposed, a state-of-the-art model in clavicle and ribs segmentation, presenting a U-Net base model with feature adaptation skip layers. Regrettably, this document's model is trained over a proprietary dataset, showing excellent, but incomparable, results. In [30] another approach for rib segmentation is taken, using Mask-RCNN models, with lower results than the presented by [29], but hanging on the same data-incomparability principle.

Therefore in this work we evaluate different attention strategies over a U-Net architecture for chest X-ray segmentation. We evaluate our proposal for lung, heart, clavicles and ribs segmentation.

2.3. Quality Assurance Models

End-to-end models[5] have been highly investigated in the last period of time. Due to this searched characteristic, quality assurance models have been mostly under-valued and under-investigated.

A rotation-determination approach [31], used generalized line histograms of rib-orientations for the accomplishment of this task. Later on, the same authors published a detailed model's evaluation in [32]. The model utilizes a SIFT-based lung segmentation algorithm to reduce

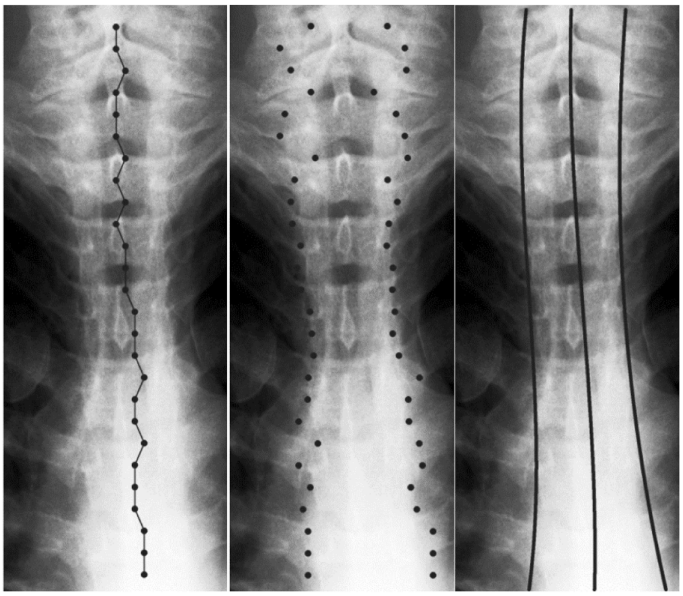
[5] End-to-end refers a model that has the capability of outputting a final result over a defined input, without the need of data pre-processing or output analysis.

the variability over the thorax's outside. This lung-segmented image is then fed to rotated gaussian filters to determine the rib's generalized line histogram, see fig. 2.4.

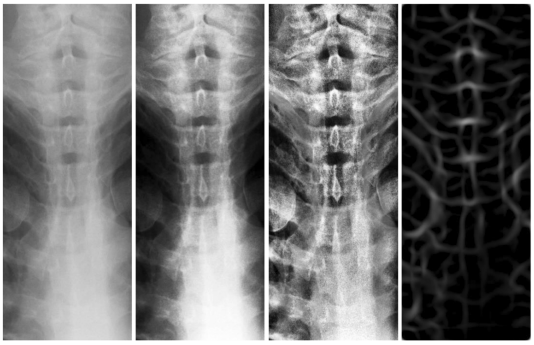
(a) Masked lungs (b) Kernels (c) Line output at 0 (d) Line output at 30° (e) Line output at 60°

Figure 2.4: Rotated Gaussian Filters Application as Shown in [32]

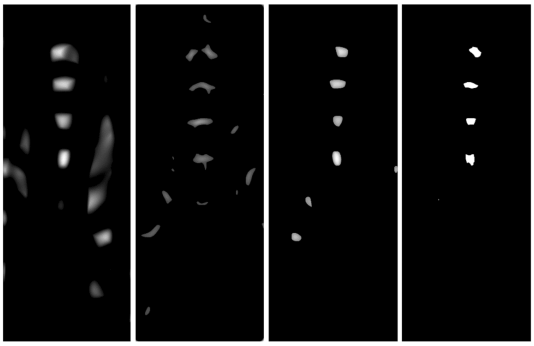
For radiation penetration, an image hardness-based approach was presented in [33]. The model is based over a pretrained ResNet-34 to determine the spine area, and a modified ridge chart structure to determine the vertebraic boundaries for counting, see fig. 2.5. Note that this model counts all thoracic vertebrae found, being far from the defined *vertebrae-behind-the-heart* approach convened by our associated medical team.



(a) ResNet-34 Based Spine Determination.



(b) The original Image, Standard Histogram Equalization, Adaptive Histogram Equalization and I_1 Second Ridge Structure



(c) I_2 Second Ridge Structure, Binarized I_1 Second Ridge Structure, Binarized I_2 Second Ridge Structure and Intersection of the Binarized I_1 and I_2 Second Ridge Structures

Figure 2.5: Spine Detection and Vertebrae Counting [33]

No data was found for the lung insuflation problem.

Chapter 3

U-Net Models for Thorax Segmentation

3.1. Blocks

For the investigation presented in this document, we use the U-Net base model, adding three new types of blocks.

3.1.1. Three Head Attention Block

As proposed in [25], we implement the three head attention block with the same configuration set in the cited document. This implementation allows us to study pseudo-attentional models, as well as generating variations of it and giving us a second baseline to compare to. This block is shown in figure 3.1

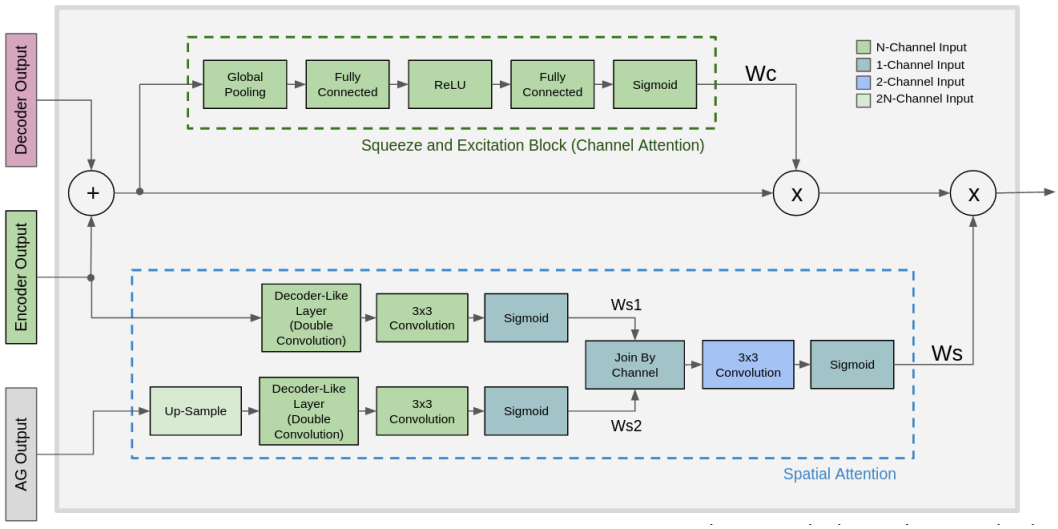


Figure 3.1: Three Terminal Attention Block as Proposed in [25]

3.1.2. Spatial Attention Block

In variation of the three head attention block, we use a simple classical image to image matrix multiplication attention in replacement to the spatial sub-block proposed in [25].

The modified block is comprised by a *Squeeze and Excitation Block* of latent size 16 and a Spatial or Full Spatial Attention Sub-Block. Two variants of the model are used, the first one presenting a single matrix multiplication of Query and Key's values, called *spatial attention*, and the second one, presenting two keys in the same multiplication, named *full spatial attention*.

$$C_k^i = (AB)_{ik} = \sum_{j=1}^{N} A_{ij} B_{jk} \tag{3.1}$$

As a matrix multiplication operates over each row and column, as shown in eq. 3.1, every position represents the relationship of the vertical and horizontal positions it represents, forgetting about the macro of the image. The concept behind a triple matrix multiplication relies behind the argument that, as the first operation connects every row and column together, while the second one considers that each pixel contains the whole column information, retrieving each row pixel in every image location. In this document, we will refer to this *triple matrix multiplication* as *full matrix multiplication*, shown in eq. 3.2.

$$D_{in} = (ABC)_{in} = \sum_{n=1}^{N} \sum_{j=1}^{N} A_{ij} B_{jk} C_{kn} \tag{3.2}$$

In figure 3.2 our proposed single matrix multiplication or *Spatial Attention* layer is shown, as described in eq. 3.1 and based over the *Three Head Attention Block* described in sec. 3.1.1.

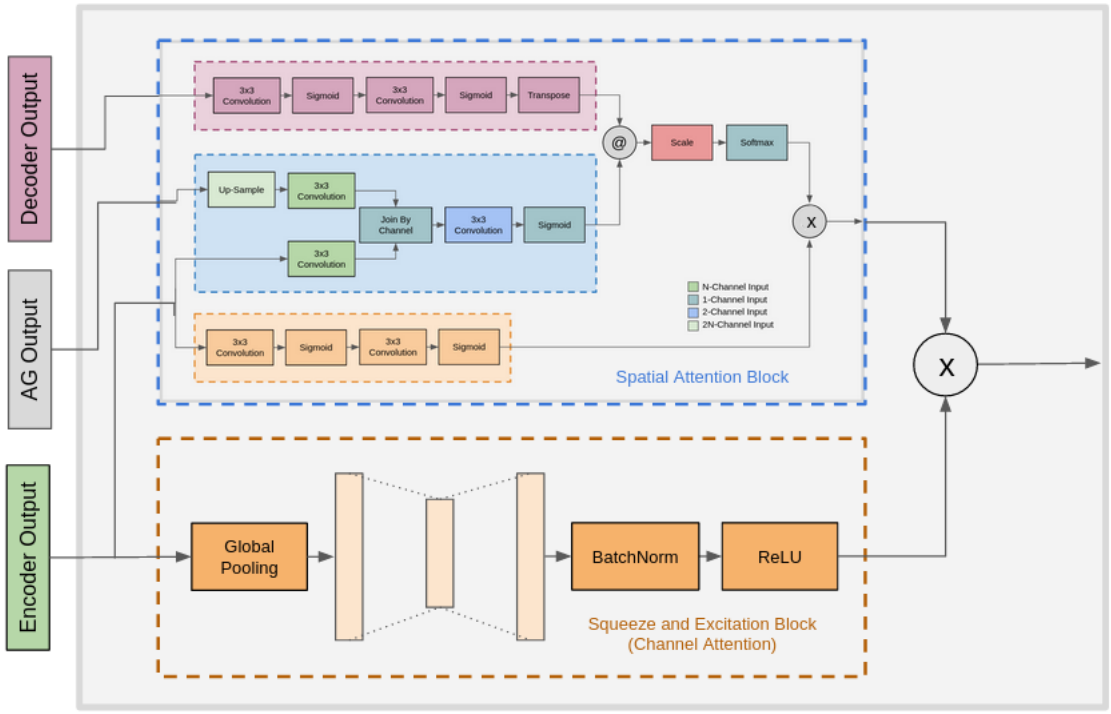


Figure 3.2: Proposed Spatial Attention Layer

As described in eq. 3.2 and based on fig. 3.2, our *Full Spatial Attention* layer is described in fig. 3.3.

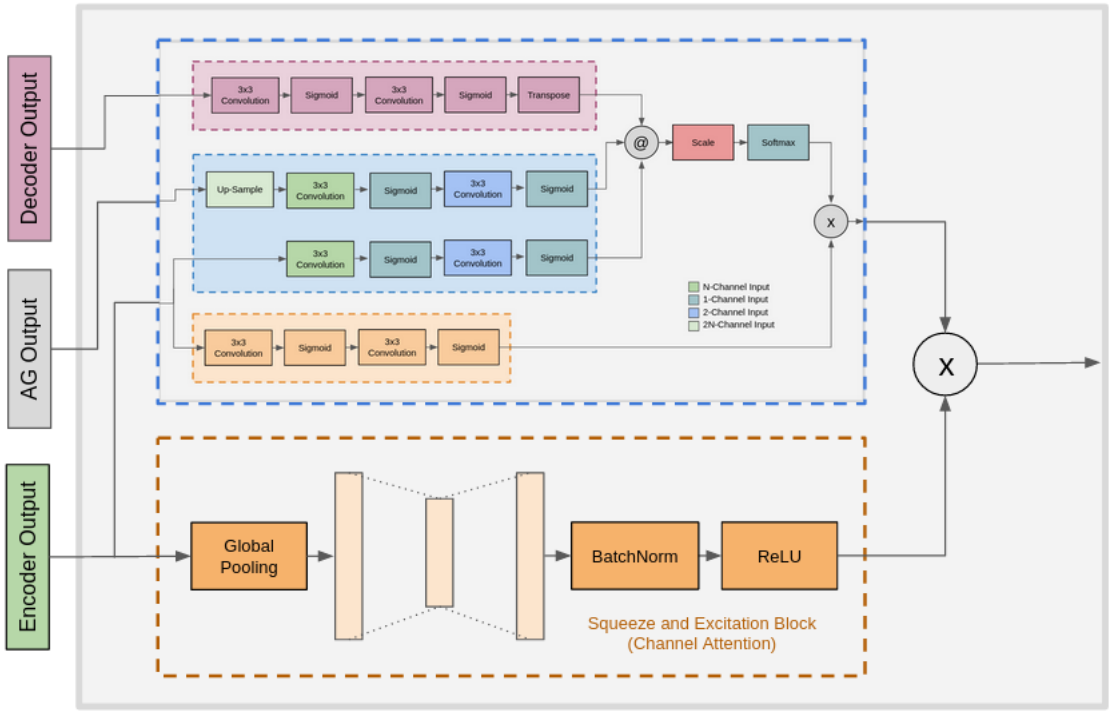


Figure 3.3: Proposed Full Spatial Attention Layer

3.1.3. Spatial Multi-Head Cross Attention Block

Based on the Spatial Attention Sub-Block presented previously, a multihead cross-attentional variation of the Spatial Attention architecture is tested. This layer-block consists in multiple concatenated Spatial Attention Sub-Blocks in parallel to an Squeeze and Excitation block as in before, but connecting only the encoder and decoder inputs as done originally in [9].

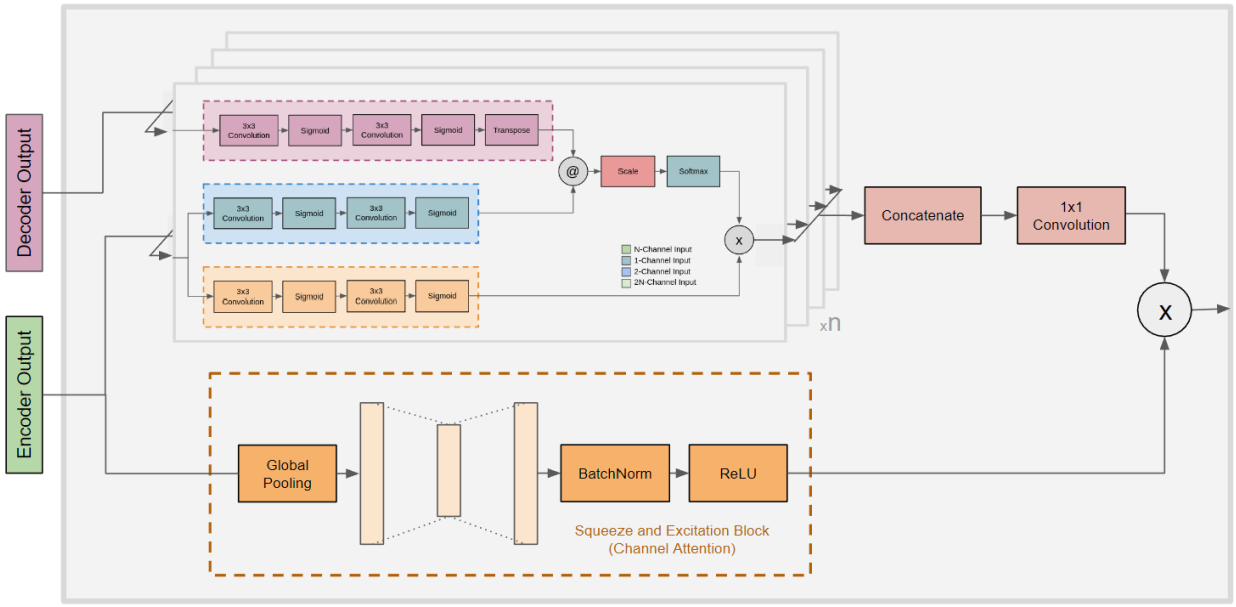


Figure 3.4: Proposed Cross-Attention Module

3.2. Architectures

In this section we will briefly define the architectures used for the development of this document. Although the descriptions are self-complete, a visual representation of each model is presented in Annex B

3.2.1. U-Net

Of the tested architectures presented, UNet serves as one of the best medical image segmentation models for baseline comparison. This architecture is used as originally presented in [11].

3.2.2. Encoder Variations

The encoder of the UNet model allows the image's characteristics extraction, so it represent an interesting reestrucuturing point to study. In this behalf, we study five UNet variations, replacing the UNet encoder with ResNet 18, 34 and 50, Swin Tranformer Base Model and a simple residual structure comprised by a single ResNet 50 [12] residual block, prior to each UNet encoder layer.

3.2.3. Skip Layer Variations

In the original UNet, the skip layers are just identity layers connecting each encoder block with it's corresponding decoder. For these experiments, we replace these layers with three-headed blocks, as presented in [25]. The blocks themselves have three inputs for the encoder, the decoder feedback and the lower skip layer. The decoder input is connected to the upsampling layer after each block and the three-headed block outputs are inputted as a normal UNet Skip layer to the corresponding decoder block. A simple representation of these connections can be seen in figure 3.5.

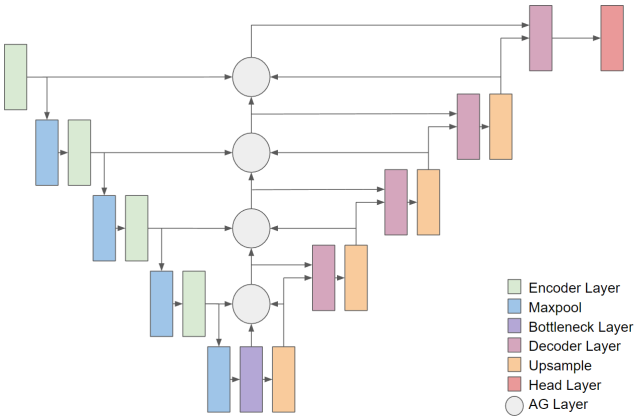


Figure 3.5: Three Head Skip Layers

3.2.4. Decoder Variations

As a proof of concept, the attention mechanism is tested at the UNet decoder. For this, we propose two variations, both adding a block prior to each decoder layer.

The first variation consists in the Cross-Attention module presented in section 3.1.3.
For the second alternative, the Spatial Attention block is modified by deleting the last layer input, leaving only one key value. This block is then furtherly modified by concatenating the encoder and decoder outputs and giving them as a fully self-attention layer [9], shown in fig. 3.6.

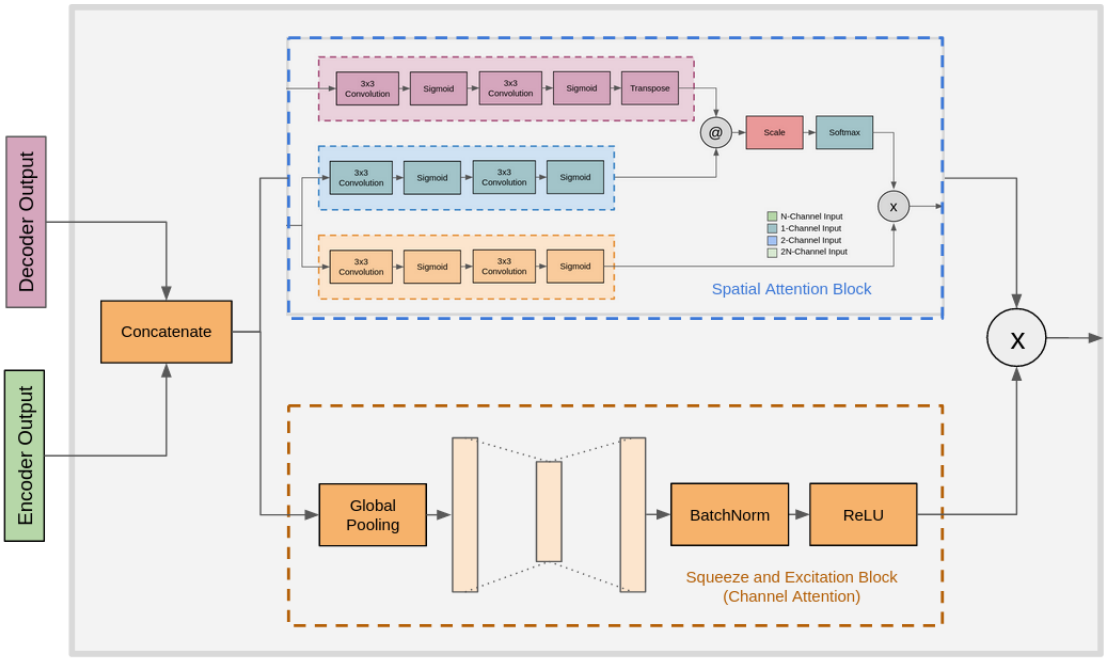


Figure 3.6: Proposed Modified Spatial Attention Block

Chapter 4

Data and Image Preprocessing for Thorax Segmentation in U-Net Models

4.1. Datasets

For these experiments, we used four public zone-specific semantic segmentation datasets: JSRT Dataset for clavicles and heart, Montgomery County Tuberculosis Dataset, Shenzhen Tuberculosis Dataset for lung segmentation, and VinDr-RibCXR for rib segmentation.

4.1.1. JSRT Dataset

This dataset includes 247 12-bit grayscale raw frontal chest X-ray images from the Japanese Society of Radiological Technology. It comprises 154 nodule-presenting images; including 100 malignant and 54 benign; and 93 normal cases. It provides basic patient report information (age and gender), nodule type diagnosis, nodule center coordinates and a basic nodule location map. The SCR dataset is a JSRT complementary set that adds full size segmentation masks for the images within the original database. It includes the independent, left and right, masks for clavicles and lungs, and the heart segmentation mask for the JSRT images.

4.1.2. Montgomery County TB Dataset

This dataset includes 138 12-bit grayscale frontal chest X-ray images from Montgomery County's Tuberculosis screening program [34]. There are 58 images with the presence of tuberculosis and 80 images free of tuberculosis. The dataset includes primary patient reports and lung segmentation masks, including age and gender.

4.1.3. Shenzhen TB Dataset

This dataset includes 662 frontal grayscale chest X-rays from Shenzhen's No.3 People's Hospital [34]. It comprises 336 cases with tuberculosis and 326 tuberculosis-free cases, including pediatric antero-posterior (AP) images. As the Montgomery Dataset, it provides basic patient reports (age and gender) and lung segmentation masks.

4.1.4. VinDr-RibCXR Dataset

This dataset is a private on-demand image set for automatic segmentation and labeling of individual ribs from chest X-ray (CXR) scans [35]. The VinDr-RibCXR contains 245 CXRs with corresponding ground truth annotations provided by human experts. All scans have been de-identified to protect patient privacy. Each image was assigned to an expert, who manually segmented and annotated each of 20 ribs, denoted as L1→L10 (left ribs) and R1→R10 (right ribs). The masks of ribs were then stored in a JSON file that can later be used for training instance segmentation models.

4.1.5. PadChest

This dataset includes more than 160,000 chest x-rays from 67,000 patients [36]. It presents 192 different classes for image multi-label classification, from which 39,039 are manually labeled and 121,829 are machine classified. For this document, we use a 26,387 image sub set of manually adults-only frontal chest x-rays, excluding the pediatric, automatically classified and lateral images of the database.

4.2. Image Preprocessing

As medical images represent space-sensitive information, image preprocessing for these tasks is greatly reduced. In this section, the input image preprocessing is described.

4.2.1. Resizing

For U-Net models, the standard 512x512 input size is used. This size allows the model to capture detailed information, without overloading the model's capabilities. This is important for attentional modules, because they scale quickly given the input size.

4.2.2. Contrast and Histogram Equalization

Grey-scale images, such as chest x-rays, lack of hue and saturation, being described mostly by its contrast. Histogram equalization methods were developed to enhance image's contrast, what could be beneficial to x-ray analysis.
An experiment we developed over the base U-Net model, showed that general histogram equalization improves as much as 5.5 additional points over segmentation results, depending of the segmented structure. For CLAHE[6], results improved only for bone structures, showing worse results for lesser dense ones, being surpassed by traditional histogram equalization in all tests. A representation of each of the two methods is shown in figure 4.1.

Given this small analysis, classical histogram equalization is a step taken during evary image's preprocessing.

4.2.3. Data Augmentation

Due to the sensitivity of medical image data, data augmentation techniques were not used for these experiments, as they represent possible medical conditions that are out of the boundaries of ambulatory medicine.

[6] CLAHE is an adaptive histogram equalization method that enhances local contrast instead of the general histogram.

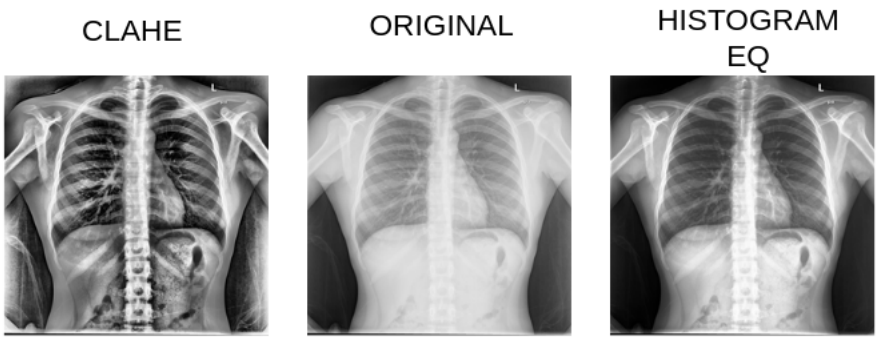


Figure 4.1: Histogram Equalization Methods

Chapter 5

Supervised Training Methods for U-Net Models in Thorax Segmentation

5.1. Training Methodology

For training these networks we use PyTorch over a single RTX3080 with batch size 8 and RMSProp optimizer with base learning rate of $1e-5$ and ReduceLROnPlateau with factor 0.5 and patience 2 for scheduler. For the training losses, we use Cross-Entropy and DICE losses, summed together. These metrics and optimizers were extensively tested over the base U-Net model, showing the best overall performance over the available PyTorch optimizers. The scheduler is set every epoch over the validation DICE Score.

For each variation, the performance shown corresponds to random weights initialization and a pretrained model over the Lungs dataset. It is also included, as completion matter, the classification-pretrained encoder version for the *Encoder Variations* models, as the other models over classification-pretrained encoders doesn't present any alteration in the overall results.

5.1.1. Evaluation

Each model has been trained 3 times, presenting the average DICE Score performance and IoU estimation over this value [$IoU = \frac{DSC}{2-DSC}$].

5.2. Result Analysis

In this section we present the results, divided by each specific variation.

5.2.1. Encoder Variations

Each encoder variation was trained over a randomly initialized set of weights and a lung-pretrained one. Every encoder was previously pretrained on ImageNet 1k. The next sections show the results given by each training variation.

As seen in table 5.1, the encoder variations improve consistently over the Montgomery-Shenzhen lungs dataset, but mostly fail in the lower image-count sets. These models decrease

Table 5.1: Encoder Variations Over Randomly Initialized Weights

Model	Montgomery-Shenzhen Lungs	VinDr Rib-CXR Ribs	JSRT Heart	JSRT Clavicles
Images (Train/Val)	610/68	221/24	222/24	222/24
U-Net	***0.930***	***0.876***	***0.798***	***0.481***
Residual U-Net	***0.934***	***0.883***	***0.831***	***0.511***
ResNet-UNet-18	***0.956***	***0.864***	***0.812***	***0.525***
ResNet-UNet-34	***0.953***	***0.862***	***0.815***	***0.515***
ResNet-UNet-50	***0.949***	***0.859***	***0.815***	***0.522***
Swin-UNet	***0.933***	***0.782***	***0.768***	***0.469***

their performance proportionally to the encoder's depth in residual backbones, result probably due to insufficient training data for these model's depths.
For the fully attentional encoder model, Swin-UNet, the 0.3 point improvement over the base U-Net model shows considerably lower results over the nineteen-point improvement given by the ResNet-UNet-34 model.
For the lower image-count datasets, the performance stands out over the shallower models, revealing a lack of training data for these heavy-weighted architectures. By this same consideration, we do not yet discard the Swin-UNet model for further studying in larger image-count datasets.

Table 5.2: Encoder Variations Over Lung-Pretrained Weights

Model	VinDr Rib-CXR Ribs	JSRT Heart	JSRT Clavicles
Images (Train/Val)	221/24	222/24	222/24
U-Net	***0.870***	***0.805***	***0.521***
Residual U-Net	***0.872***	***0.818***	***0.524***
ResNet-UNet-18	***0.865***	***0.815***	***0.532***
ResNet-UNet-34	***0.857***	***0.813***	***0.524***
ResNet-UNet-50	***0.856***	***0.801***	***0.520***
Swin-UNet	***0.788***	***0.801***	***0.480***

In 5.2, the lung-pretrained models show a considerably better begining during the training procedure, presenting results over 80% of the best validation DSC score, on epoch one. Regretfully, these models confirm the lack of training data, by overfitting around epoch two, presenting almost the same results than the non-pretrained models, but in extremely lower training epoch count.

Finally, the classification-pretrained encoders over these models in figure 5.3, show little to none improvement over the non-pretrained ones, possibly due to task-difference from the first layers taken by the skip connections. This results propose the study over an hybrid model between enhanced skip layers, for processing and transforming the raw classification

Table 5.3: Encoder Variations Over Classification-Pretrained Weights

Model	Montgomery-Shenzhen Lungs	VinDr Rib-CXR Ribs	JSRT Heart	JSRT Clavicles
Images (Train/Val)	610/68	221/24	222/24	222/24
U-Net	***0.930***	***0.876***	***0.798***	***0.481***
ResNet-UNet-18	***0.950***	***0.864***	***0.792***	***0.532***
ResNet-UNet-34	***0.951***	***0.853***	***0.785***	***0.501***
ResNet-UNet-50	***0.946***	***0.864***	***0.796***	***0.498***
Swin-UNet	***0.817***	***0.701***	***0.813***	***0.479***

data from the encoder to the characteristic-extracted decoder input.

5.2.2. Skip Layer Variations

For the skip layer variations, the same procedure of the encoder variation's section is used. Due to the classification-pretrained non-variant results over the base models, this data is not shown, as it is almost identical to the results shown in table 5.4.

Table 5.4: Skip-Layer Variations Over Randomly Initialized Weights

Model	Montgomery-Shenzhen Lungs	VinDr Rib-CXR Ribs	JSRT Heart	JSRT Clavicles
Images (Train/Val)	610/68	221/24	222/24	222/24
U-Net	***0.930***	***0.876***	***0.798***	***0.481***
Three-Head Attention U-Net	***0.948***	***0.880***	***0.812***	***0.485***
Spatial Attention U-Net	***0.959***	***0.870***	***0.807***	***0.475***
Double Spatial Attention U-Net	***0.946***	***0.863***	***0.803***	***0.493***
Full Spatial Attention U-Net	***0.959***	***0.872***	***0.817***	***0.489***
Swin Spatial Attention U-Net	***0.926***	***0.878***	***0.808***	***0.513***

As presented in table 5.4, attentional skip layers show an improvement over the base U-Net architecture due to their aggregated complexity and learnable parameters. For this same reason, saying that attentional skip layers are better than every other block would be a rushed conclusion.

As seen in table 5.5, lung pre-training does not show great variations on the segmentation dice score, lowering, in most cases, the final score of each model. This can be described by attention and pseudo-attentional mechanisms used in this category. Attentional models search for area-specific information. As such, the lungs take almost the whole thorax region, being spatially distant in some space-specific areas.

5.2.3. Decoder Variations

Decoder variations affect the way the information is mixed and the image-size object rebuilt. As it is dependant on the characteristics extraction given by the encoder, and the feature

Table 5.5: Skip-Layer Variations Over Lung-Pretrained Weights

Model	VinDr Rib-CXR Ribs	JSRT Heart	JSRT Clavicles
Images (Train/Val)	221/24	222/24	222/24
U-Net	***0.870***	***0.805***	***0.521***
Three-Head Attention U-Net	***0.873***	***0.799***	***0.538***
Spatial Attention U-Net	***0.875***	***0.809***	***0.545***
Double Spatial Attention U-Net	***0.858***	***0.799***	***0.481***
Full Spatial Attention U-Net	***0.865***	***0.820***	***0.503***
Swin Spatial Attention U-Net	***0.868***	***0.800***	***0.473***

adaptation of the skip layers, it is not expected to improve in detail finding, but working better in the output generation quality and pixel details.

Table 5.6: Decoder Variations Over Randomly Initialized Weights

Model	Montgomery-Shenzhen Lungs	VinDr Rib-CXR Ribs	JSRT Heart	JSRT Clavicles
Images (Train/Val)	610/68	221/24	222/24	222/24
U-Net	***0.930***	***0.876***	***0.798***	***0.481***
Spatial Decoder	***0.800***	***0.851***	***0.810***	***0.538***
Cross Attention	***0.820***	***0.853***	***0.775***	***0.415***

For U-Net models over non-pretrained weights, as seen in table 5.6, attentional-complexity layers show a reduction in the model learning capabilities and lower DICE Scores.
Attention layers are mostly used for area-specific semantic depictions or time-space variant characteristic extractions. This is far from the concept over the U-Net model's decoder, which reconstructs and upsamples the retrieved information.

Table 5.7: Decoder Variations Over Lung-Pretrained Weights

Model	VinDr Rib-CXR Ribs	JSRT Heart	JSRT Clavicles
Images (Train/Val)	221/24	222/24	222/24
U-Net	***0.870***	***0.805***	***0.521***
Spatial Decoder	***0.862***	***0.818***	***0.517***
Cross Attention	***0.828***	***0.770***	***0.494***

As well as presented in previous cases, lung pretraining does not show a big variation over non-pretrained models. The low image-count need of the models is compensated completely after a couple of epochs, showing similar results to pretrained models. Even though pretrain

does not considerably affect the output, it reduces extremely the time needed to train the model.

Chapter 6

Quality Assurance Determination Based on Semantic Chest Structures Segmentation

For determining an image's quality assurance, it is important to note the segmentation capabilities that we currently have. For accomplishing this objective, we have to consider a baseline clavicle segmentation, a relatively good heart and rib segmentation, a very well made lung segmentation and no vertebrae segmentation.

6.1. Centered Image and Segmentation Pre-Check

During the evaluation of the models presented, a lot of *Non-Centered* images and black-masks were encountered. To tackle this problem, we use the lung segmentation mask to determine if the image can be considered *centered* and the image pixel-sum to determine if there is an actual, an viable, segmentation detected.

6.2. Patient Rotation

As described in section 1.2.1.2, the patient's rotation is determined by the clavicle's head to the medial line. For approaching this problem, we use the lung segmentation as a mean to find the medial line, as the lung horizontal borders are symmetric regarding the medial plane. The clavicle horizontal margins are determined by a simple axial sum on the horizontal plane. The detailed procedure is as it follows:

1. Lung Border Determination
    a) Extract the best lung segmentation from the image.
    b) Apply a 5x5 kernel-size Gaussian Blur with standard deviation 5 to reduce small pixel outliers in the lung's outer section.
    c) Sum the mask's pixels vertically to get a x-axis vector representation histogram.
    d) As the lung segmentation results are highly detailed, we are confident to determine the lung margins as the leftmost and rightmost non-zero value.
    e) The medial line is determined as the mean value of the lung's border limits. A simple visualization of these margins is shown in fig. 6.2.a.

2. Clavicle Horizontal Margins Determination

   a) Extract the best possible clavicle segmentation mask from the input image, as shown in fig. 6.1.a.

   b) Sum the mask's pixels vertically to get a x-axis vector representation histogram, see fig. 6.1.b.

   c) As the clavicle representation isn't as robust as lungs segmentation, we binarize the x-axis histogram. In this step is important to note that the clavicle's segmentation could be non-contiguous, being composed by a set of smaller blocks instead of two independent full ones, see fig. 6.1.c.

   d) To determine the central boundaries of a composite clavicle, we multiply the binarized histogram vector with a medial-line-centered gaussian distribution, resulting in sector-defined split gaussian vector, see fig. 6.1.d.

   e) The central borders are selected as the maximum argument of the split distribution for each side of the medial line.

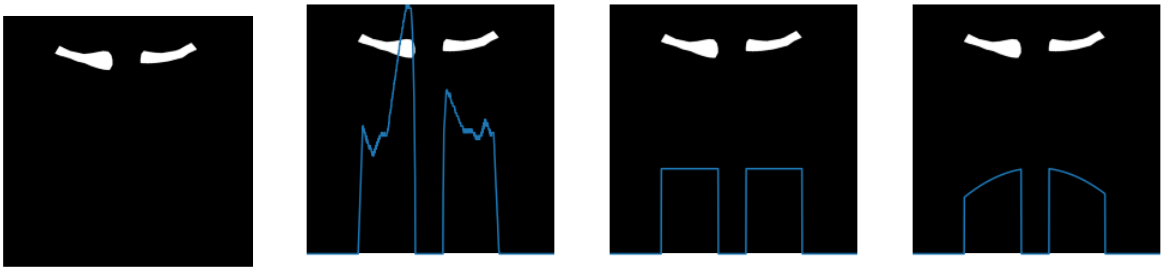
   f) To determine the border boundaries of this composite clavicles, a similar approach is taken with an *upside-down inverted* gaussian distribution.

   g) The extreme borders are selected the same way as central limits: as the maximum argument of the *inverted* split distribution for each side of the medial line.

3. Rotation Determination

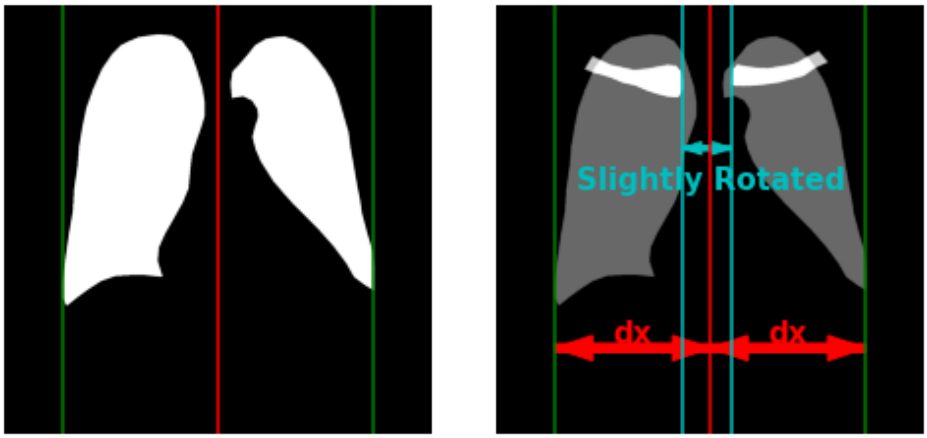
   a) Taking in acquaintance the clavicle margins and the medial line, we determine the rotation as the relation of the clavicle's boundaries with the medial line, as shown in figure 6.2.b.

   b) With a threshold tolerance of 50%, we consider an image *Rotated*. For the [15%, 50%] range, the image is considered as *Slightly Rotated*. Finally, the values lower than 15% variation, are considered *Correctly Positioned*.



(a) Step 2.a (b) Step 2.b (c) Step 2.c (d) Step 2.d

Figure 6.1: Step 2: Clavicle Border Determination



(a) Lung Borders and Medial Line (b) Rotation Determination

Figure 6.2: Rotation and Medial Line Determination

6.3. Lung Insuflation

As described in section 1.2.1.3, the lung insuflation is determined by counting anterior and posterior ribs. As our segmentation method does not separe anterior from posterior ribs, we approach the problem as the count of superposed rib count. As they are 12 posterior ribs and 10 anterior ribs, we hope to see 10 and 6 of them, respectively. Depending the patients height, some rib cages have more separation between ribs than other people. For this case, we found that using a superposed rib count in the range]7,9] led to best insuflation results. The rib superposition count is done as follows:

1. As described by [32], we apply rotated gaussian kernels. These kernels are determines as follows:
   a) Define a kernel-size zero array.
   b) Set every horizontal-center pixel to a value of 1, to determine the rotation angle.
   c) Symmetrically set each pixel to a value of a defined alpha to the power of the distance to the farthest kernel border to expand the line width. For this definition, the value of alpha must be in the range [0,1].
   d) Rotate this kernel to the desired angle.
   e) Normalize the kernel, so the sum of pixels be 1.
   f) Multiply the kernel to a 2D gaussian distribution of standard deviation sigma.
2. Determine the horizontal line histogram.
   a) With a previously calculated rotated kernel[7], we filter the input image. Steps 2.a, 2.c and 3.g are shown in a composite visual representation presented in figs. 6.3.c and 6.3.f.

[7] Kernel properties: size 121x121, rotation 180°, alpha 0, sigma 51

b) To obtain side-by side the vertical line histograms, we divide the image in half by the medial plane. As the image already has been tested to be centered, we ensure that each section contains, exactly, half of the thoracic cage.

c) We sum, vertically, the pixels values of each of the halves of the filtered image to obtain the histograms. Steps 2.a, 2.c and 3.g are shown in a composite visual representation presented in figs. 6.3.c and 6.3.f.

3. For counting the superposed ribs, we apply a centroid-based algorithm as follows:

a) Determine the non-zero index values of the array. Note that the used histogram is one-dimensional.

b) Obtain the first derivatives of the array. For this determination, we substract each index value with the next and vice-versa, obtaining the two side-derivatives of the array.

c) We binarize each array, with value 0 corresponding to each 1-pixel difference (eg. 156, 157 -> 158-157 = 1), and value 1 to all others. It's important to note these last two steps as a determination of the greater-than-zero pixel-set borders.

d) We define each set of borders as the maximum and minimum index value of each greater-than-zero section.

e) We propose each of the defined sections to be centroid candidates.

f) We repeat these steps for the other half of the filtered image.

g) We determine the rib-count as the mean value of total candidates. Steps 2.a, 2.c and 3.g are shown in a composite visual representation presented in figs. 6.3.c and 6.3.f.

(a) Original Image (b) Lung Segmentation (c) Steps 2.a, 2.c and Centroid Classification

(d) Original Image (e) Lung Segmentation (f) Steps 2.a, 2.c and Centroid Classification

Figure 6.3: Superposed Rib Clusterization Counting. On top: a complex case where ribs are extremely superposed, on bottom: a standard superposition case.

6.4. Radiation Penetration

Due to the lack of spinal-region or vertebrae segmentation, the penetration determination is the hardest of them all. For this section we will be detailing a behind-the-heart vertebrae count determination based on [33]. For this counting, we use only the heart shape mask and the original image, determining the vertebrae count by image-processing means.

1. Image Pre-Processing

   a) Given the base image and it's heart segmentation map, we redefine the image as the multiplication of them both, leaving only heart-masked area. For better detection, the image is grayscale-inverted and histogram-equalized before masking.

   b) For determining the heart position, vertical and horizontal pixe-wise additions are used. For preventing small-sized outliers, a 5x5-size kernel gaussian blur is applied. The vertical an horizontal pixel sum is visually represented in fig. 6.4.a.

   c) As presented in section 1.2.1.2 a simple binarization of each axis's sum is applied.

This binarization allows us to have a simple step function for both axes, as shown in fig. 6.4.b.

d) The maximum and minimum arguments are determined by the index of the first and last non-zero values over the vertical and horizontal sums.

e) Given the limits of the heart area, the image is cropped to show the minimum rectangle containing the full heart's shade, as shown in figure 6.4.c.

2. Spine-Area Detection

a) For determining the spine area, bests results are got using the application of CLAHE[8] over the image, as shown in fig. 6.4.d.

b) After the CLAHE application, a uniform-like filter is applied to the image. The filter kernel corresponds to a 3x3 ones matrix, with central value of -8, assuring the total sum of values corresponds to 0.

c) For smoothing the detected rough borders, a 3x3 size gaussian blur kernel is applied to the image.

d) As the spine represents the brightest area in the heart's shadow, an x-axis sum is applied.

e) The spine boundaries are determined by the 25% brightest contiguous horizontal pixel areas in the image.

f) We return a cropped version of the spine area, see fig. 6.5.a.

3. Vertebrae Detection

a) For best results, we observed that applying a 7x3-kernel gaussian blur works better. This stage horizontally expands the bright areas and reduces the small-size outliers in the vertical axis.

b) As presented in [33], we determine the four second derivatives of the image, named $L_{xx} = \frac{\partial^2}{\partial x^2},\ L_{xy} = \frac{\partial^2}{\partial x \partial y},\ L_{xx} = \frac{\partial^2}{\partial y \partial x},\ L_{xx} = \frac{\partial^2}{\partial y^2}$.

c) We reshape each of the L tensors to a 1 dimensional array and concatenate them pixel-wise in a matrix $L = \begin{pmatrix} L_{xx} & L_{xy} \\ L_{yx} & L_{yy} \end{pmatrix}$

d) For each dimension of obtained tensor, we derive the eigenvalues of the 2x2 matrix. Note that these eigenvalues are all real-type data.

e) For each of the two eigenvalues for each matrix, we construct two new vectors with the maximum dimension-values and minimum dimension-values, respectively.

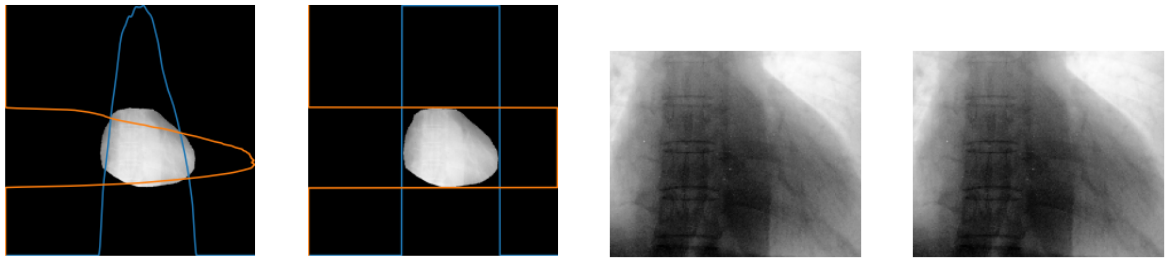
f) The maximum and minimum vectors have the same size as the original spine area, and they are reshaped to their original form. Now we have two images composed by pixel-wise maximum and minimum eigenvalues of the second derivative of the spinal area. These maximum and minimum second-derivative eigenvalues are represented in figs. 6.5.b and 6.5.c, respectively.

g) With threshold values 20 and 10, we binarize the maximum and minimum eigen-images, as shown in figs. 6.5.d and 6.5.e.

[8] CLAHE is an adaptive local contrast adjusting method.

h) The intersection of these images is determined by multiplicating their binary masks, see fig. 6.5.f.

i) Finally, the vertebrae are counted the same way presented in section 1.2.1.3, step 3 (centroid-based counting). This centroid-determined number corresponds to the number of visible vertebrae in the spine defined area.



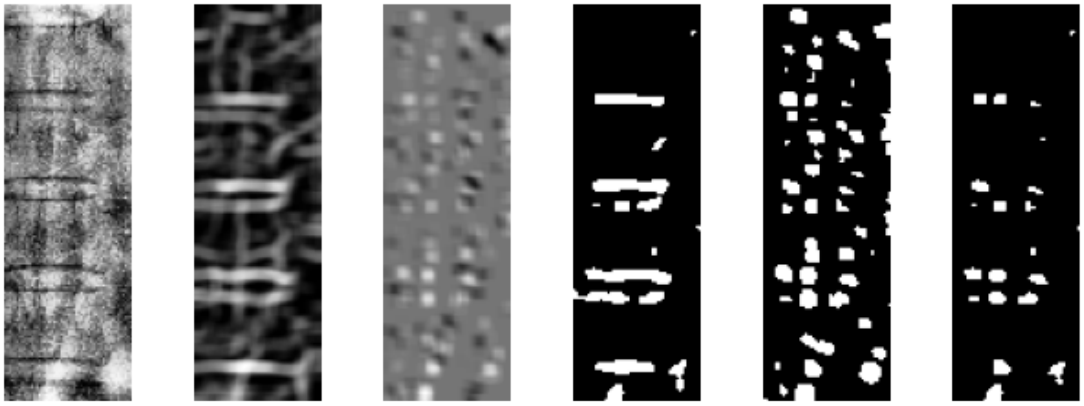
(a) Step 1.b Vertical and Horizontal Pixel-Sums

(b) Step 1.c Binarized Vertical and Horizontal Pixel-Sums

(c) Step 1.e Cropped Heart Shade

(d) Step 2.a Cropped CLAHE-Applied Heart Shade

Figure 6.4: Heart Boundaries Determination



(a) CLAHE-Applied Spine Area

(b) Step 3.f Maximum Eigenvalues Visualization

(c) Step 3.f Minimum Eigenvalues Visualization

(d) Step 3.g Binarized Maximum Eigenvalues Visualization

(e) Step 3.g Binarized Minimum Eigenvalues Visualization

(f) Step 3.h Visualization

Figure 6.5: Step 3: Vertebrae Counting Steps Visualization Based on [33]

6.5. Result Analysis

Due to the lack of quality assurance datasets, a systematic approach to determining the validity was taken. For accomplishing this objective, our associated radiologists selected an

image subset with well distributed data over 8 multi-label classes, being:

1. Penetration Classes
    a) Under-Penetrated
    b) Correctly-Penetrated
    c) Over-Penetrated
2. Insuflation Classes
    a) Under-Insuflated
    b) Correctly-Insuflated
    c) Over-Insuflated
3. Rotation Classes
    a) Non-Rotated
    b) Rotated

It is important to note that a correctly taken normal image corresponds to the classes: Correctly-Penetrated, Correctly-Insuflated and Non-Rotated. 50 randomly selected chest x-rays from this subset were used for determining the confusion matrices and precision, f1-score, recall and accuracy values for each class.

As there is almost no data for radiological image quality determination, we will approach the problem by using standard classification metrics as made explicit at the beginning of this section, after the defined classes were presented. In table 6.1 we present the obtained results for f1-score, precision, recall and accuracy, over the three main class divisions. For this data, we define the true positive for correctly classified well taken images.

Table 6.1: Result Analysis For Quality Assurance Assesments

Metric	Penetration Detection	Insuflation Detection	Rotation Detection
Accuracy	0.48	0.23	0.22
Precision	0.89	0.23	0.89
Recall	0.41	1.00	0.18
F1-Score	0.57	0.37	0.30

As clavicle segmentation models are considerably weak, the bone heads are not always detected as one should expect. This causes a lot of erratic variation from the medial line, specially considering the excellent results for lung segmentation models. As we apply image processing to masks, if we deliver a good segmentation of the lungs, ignoring small-spot outliers, the medial line determination is almost perfect. Regrettably, it is not the same for the case of clavicles, as the wrongly segmented areas cause irreparable damage to the image's segmentation map, therefore, displacing the mid-point between the bone heads from

the medial plane. In rotation determination, these facts present themselves as almost no false positives, a lot of false negatives, and an almost perfect score for true negatives. This causes accuracy to be in an almost upper-central result, just above the 50% barrier, a good precision result and low recall value. The confusion matrix for general rotation is shown in fig. 6.6.

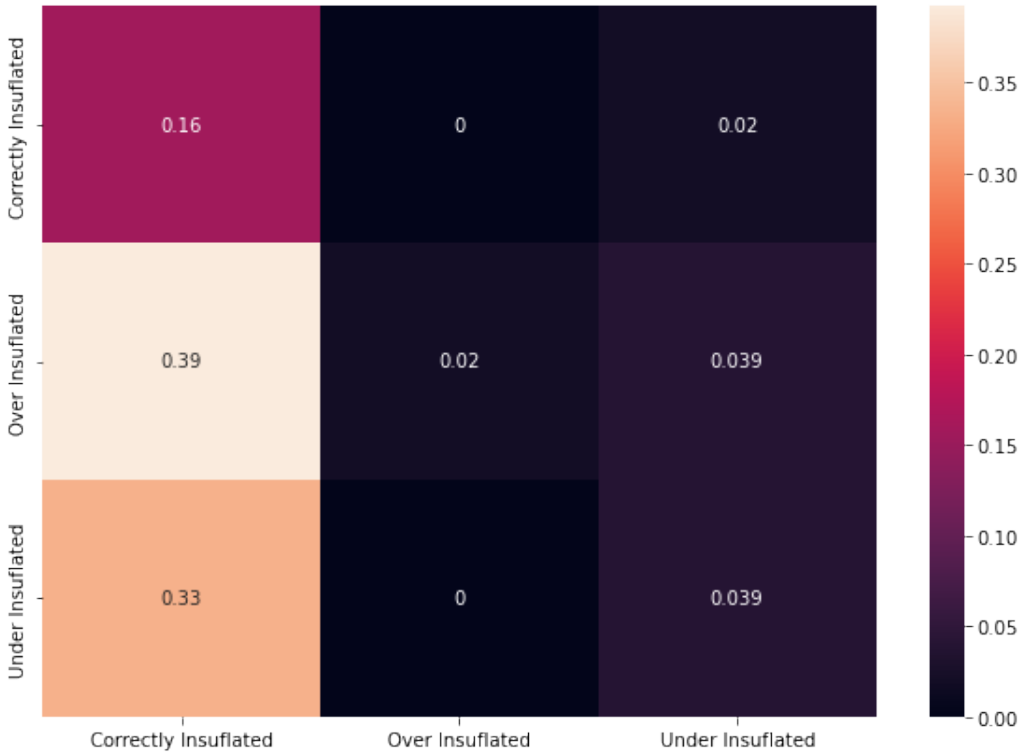


Figure 6.6: Rotation Confusion Matrix

For the insuflation problem, the rib segmentation has shown good results, but, as shown in fig. 6.3, the superposition of posterior ribs sometimes is excessive, leading to a higher centroid-count. This problem is visible in the augmented false negatives count, lowering the result of true positive values. As the rib superposition is augmented on over-insuflated cases, the centroid count is also augmented, presenting lower results for the under-insuflated detection. This under-insuflated cases are commonly detected as correctly-insuflated cases, which leads to an augment in false positives for the mentioned case, but showing good results in true negatives detection for the general insuflation class. These results lead to the decreased precision score, and almost perfect recall value. The confusion matrix for general insuflation is shown in fig. 6.7.

Finally, for the penetration case, the thing is a little more complex, due to the third image processing step. Using thresholds to define certain conditions can be, sometimes, misleading. As the contrast-ratio of different images vary, threshold values tend to work better in some cases than others. In this case of application, it is important to note that the image is first histogram-equalized, and re-equalized with CLAHE support for detection, leading to a more standard, yet not perfect, contrast balance. Heart segmentation, on the other hand, does not show perfect results, sometimes enlarging the heart's shade and, for instance, augmenting the vertebrae-count, leading to a greater over-penetration classification at true positives cost.

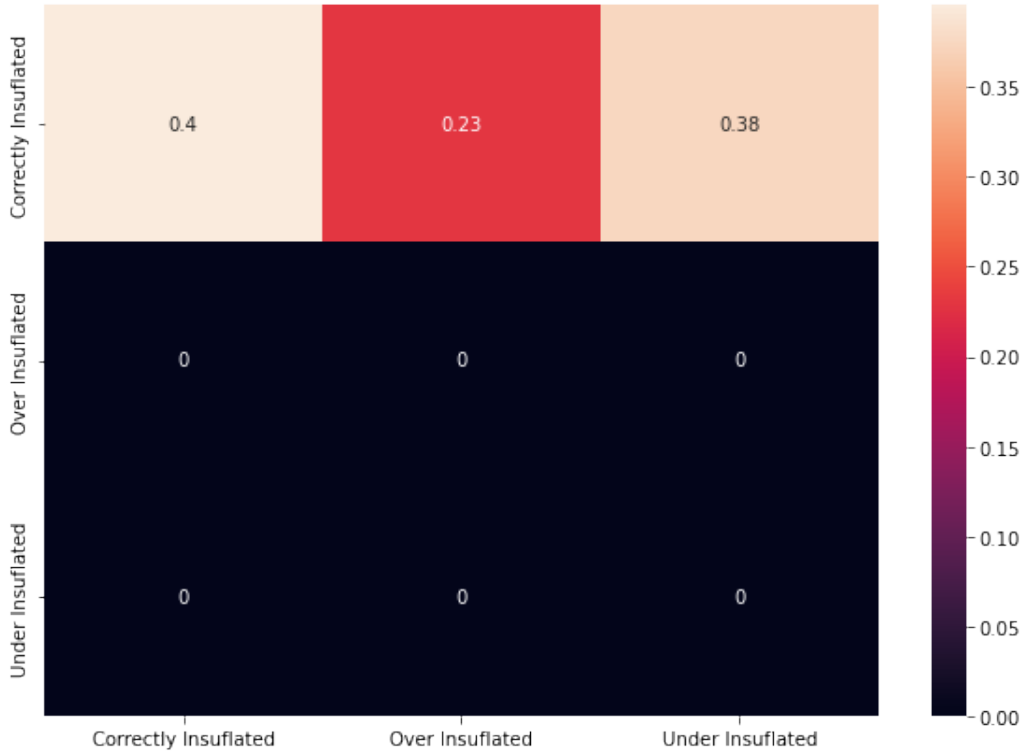


Figure 6.7: Insuflation Confusion Matrix

As seen in figure 6.2, thresholding and filtering not always removes small outliers, which results in higher vertebrae count, supporting the false negative augment presented in the last point. These results show diminished results in all metrics, being f1-score and recall the lowest of them all. The confusion matrix for general penetration is shown in fig. 6.8.

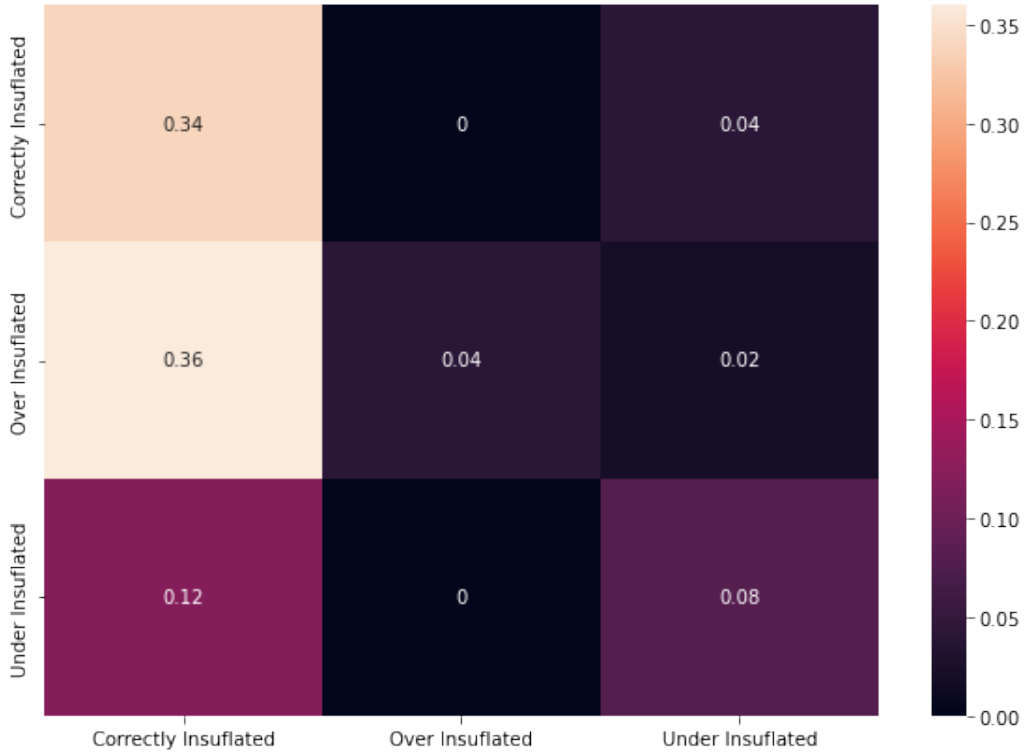


Figure 6.8: Penetration Confusion Matrix

Chapter 7

Conclusions and Discussion

7.1. Discussion

7.1.1. U-Net Based Chest X-Ray Image Segmentation

Semantic segmentation is one of the most complex computer vision problems, being considered AI-hard, the same level as human emotion-comprehensive systems. As well as segmentation is a complex area, the number of manually segmented images in medical areas are hard to find.
Considering the points listed above, medical image segmentation is an ongoing investigation work, which is far from perfect.
Observing the obtained results in table 7.1, our models are near state-of-the-art results.

Table 7.1: Dice Score Comparison over Used Datasets

Model	Lung Segmentation	Rib Segmentation	Clavicle Segmentation
Images (Train/Val)	221/24	222/24	222/24
[26]	***0.9740***	-	-
[27]	***0.8503***	-	-
[22]	***0.9496***	-	-
[25]	***0.9584***	-	-
[28]	***0.9483***	-	-
[24]	***0.9796***	-	-
[30][1]	-	***0.7340***	-
[29][1]	-	***0.8838***	***0.9378***
Our [2]	***0.9590***	***0.8780***	***0.5380***

As most data scientists and machine learning engineers would like more data, medical

[1] Use of proprietary training dataset which does not compare to the training conditions shown in our work.
[2] This shows our best models for each task. For lungs, it represents *Spatial Attention U-Net* model. For Ribs, it represents the *Swin-Block Skip* architecture. For clavicles, it represents the *Three-Head Attention U-Net* model

image segmentation and classification is a high-cost ethics-protected work, that will barely be published, fomenting the development of models in need for lesser image counts, being further apart from modern image segmentation architectures.

Semi-supervised learning has been one of the raising investigation areas that, hopefully, will permit medical models to be fully viable, thus it will not replace the basic image segmentation models and lower image-count models due to the necessity of fine-tuning data. This way, further models proposed should be structured and based on semi-supervised learning over low image-count needy architectures.

Recent open-source work in x-ray analysis [37] and [38] present trained models and multiple techniques for image classification and 3D image segmentation.

7.1.2. Quality Assurance Determination for Chest X-Ray Imaging

Even though quality assurance is barely mentioned in the existing bibliography due to the classical end-to-end approaching attempts, it adds multiple image capabilities, ensuring a good quality over out-of-the-box imaging.
The patient's rotation determination is approached in [32] with quite unfruitful results. The same was done in [33] for penetration determination, but with no ideal results.

7.2. Ongoing Work

By the moment of the delivery of this document, a paper associated to the segmentation of chest x-ray images is being reviewed for publication, addressing the themes presented in chapters three trough five. This paper will be published under the name "Impact of Attention Modules for Organ Segmentation in Chest X-ray Images under U-Net Architectures" (de la Sotta *et al.*, 2022) for further reading, not yet published.
Two digital posters are available at the moment, presented in the *European Congress of Radiology 2023*[9], entitled "Attentional Layers Improve U-Nets Multi Structure Chest X-Ray Segmentations Using Less Training Data" (ECR) (de la Sotta *et al.*, 2023) and the Chilean Congress of Radiology 2022[10] (CChR) (de la Sotta *et al.*, 2022).

7.3. Future Work

Future work on this area corresponds to the study of different semi-supervised learning techniques to improve the results over the presented models. Although we tackled the rib counting and vertebrae detection in an image processing manner, a small dataset including separated anterior and posterior ribs and vertebrae is being constructed for reducing the result variability in these specific tasks.

[9] Accepted, will be published and presented during the 2023 congress, March 1-5, Viena.
[10] Published, https://congresochilenoradiologia.cl/envio-de-posters/redes-neuronales-semi-atencionales-tipo-u-net-para-segmentacion-multi-estructura-en-radiologia-de-torax/

Although *Data Augmentation* was discarded for the development of this document, the study and evaluation of new *Data Augmentation* techniques should be addressed in future work.

Bibliography

[1] Zdora, M.-C., "Principles of X-ray Imaging," en X-ray Phase-Contrast Imaging Using Near-Field Speckles, pp. 11–57, Cham: Springer International Publishing, 2021, doi:10.1007/978-3-030-66329-2_2. Series Title: Springer Theses.

[2] Jan, J. y Jan, J., Medical Image Processing, Reconstruction and Restoration: Concepts and Methods. CRC Press, 0 ed., 2005, doi:10.1201/9781420030679.

[3] Broder, J., Diagnostic Imaging for the Emergency Physician: Expert Consult. 2011.

[4] Agrawal, T. y Choudhary, P., "Segmentation and classification on chest radiography: a systematic survey," The Visual Computer, 2022, doi:10.1007/s00371-021-02352-7.

[5] Shaziya, H., Shyamala, K., y Zaheer, R., "Automatic Lung Segmentation on Thoracic CT Scans Using U-Net Convolutional Network," en 2018 International Conference on Communication and Signal Processing (ICCSP), (Chennai), pp. 0643–0647, IEEE, 2018, doi:10.1109/ICCSP.2018.8524484.

[6] Basavarajaiah, M., "6 basic things to know about Convolution," 2019, https://medium.com/@bdhuma/6-basic-things-to-know-about-convolution-daef5e1bc411.

[7] Zhang, M., Yu, Z., Wang, H., Qin, H., Zhao, W., y Liu, Y., "Automatic Digital Modulation Classification Based on Curriculum Learning," Applied Sciences, vol. 9, no. 10, p. 2171, 2019, doi:10.3390/app9102171.

[8] Leone, G., "Leyes de la Gestalt," 2018, https://guillermoleone.files.wordpress.com/2018/01/leyes-de-la-gestalt.pdf.

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., y Polosukhin, I., "Attention Is All You Need," 2017, http://arxiv.org/abs/1706.03762 (visitado el 2022-12-17). arXiv:1706.03762 [cs].

[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., y Houlsby, N., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2021, http://arxiv.org/abs/2010.11929 (visitado el 2022-12-17). arXiv:2010.11929 [cs].

[11] Ronneberger, O., Fischer, P., y Brox, T., "U-Net: Convolutional Networks for Biomedical Image Segmentation," en Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (Navab, N., Hornegger, J., Wells, W. M., y Frangi, A. F., eds.), (Cham), pp. 234–241, Springer International Publishing, 2015.

[12] He, K., Zhang, X., Ren, S., y Sun, J., "Deep Residual Learning for Image Recognition," 2015, http://arxiv.org/abs/1512.03385 (visitado el 2022-12-17). arXiv:1512.03385 [cs].

[13] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., y Guo, B., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," 2021, http:

//arxiv.org/abs/2103.14030 (visitado el 2022-12-17). arXiv:2103.14030 [cs].

[14] Srinivas, A., Lin, T.-Y., Parmar, N., Shlens, J., Abbeel, P., y Vaswani, A., "Bottleneck Transformers for Visual Recognition," CVPR 2021, http://arxiv.org/abs/2101.11605 (visitado el 2022-12-17). arXiv:2101.11605 [cs].

[15] Dice, L. R., "Measures of the Amount of Ecologic Association Between Species," Ecology, vol. 26, no. 3, pp. 297–302, 1945, doi:10.2307/1932409.

[16] Rote, G., "Computing the minimum Hausdorff distance between two point sets on a line under translation," Information Processing Letters, vol. 38, no. 3, pp. 123–127, 1991, doi:10.1016/0020-0190(91)90233-8.

[17] Zhou, K., Medical Image Recognition, Segmentation and Parsing: Machine Learning and Multiple Object Approaches (The MICCAI Society book Series). 2015.

[18] Iglesias, J. E. y Sabuncu, M. R., "Multi-atlas segmentation of biomedical images: A survey," Medical Image Analysis, vol. 24, no. 1, pp. 205–219, 2015, doi:10.1016/j.media.2015.06.012.

[19] Long, J., Shelhamer, E., y Darrell, T., "Fully Convolutional Networks for Semantic Segmentation," 2015, http://arxiv.org/abs/1411.4038 (visitado el 2022-12-17). arXiv:1411.4038 [cs].

[20] Palmer, S. E., Vision Science: Photons to Phenomenology. 1999.

[21] Liu, W., Luo, J., Yang, Y., Wang, W., Deng, J., y Yu, L., "Automatic lung segmentation in chest X-ray images using improved U-Net," Scientific Reports, vol. 12, no. 1, p. 8649, 2022, doi:10.1038/s41598-022-12743-y.

[22] Mique, E. y Malicdem, A., "Deep Residual U-Net Based Lung Image Segmentation for Lung Disease Detection," IOP Conference Series: Materials Science and Engineering, vol. 803, no. 1, p. 012004, 2020, doi:10.1088/1757-899X/803/1/012004.

[23] Charng, J., Xiao, D., Mehdizadeh, M., Attia, M. S., Arunachalam, S., Lamey, T. M., Thompson, J. A., McLaren, T. L., De Roach, J. N., Mackey, D. A., Frost, S., y Chen, F. K., "Deep learning segmentation of hyperautofluorescent fleck lesions in Stargardt disease," Scientific Reports, vol. 10, no. 1, p. 16491, 2020, doi:10.1038/s41598-020-73339-y.

[24] Gite, S., Mishra, A., y Kotecha, K., "Enhanced lung image segmentation using deep learning," Neural Computing and Applications, 2022, doi:10.1007/s00521-021-06719-8.

[25] Cao, F. y Zhao, H., "Automatic Lung Segmentation Algorithm on Chest X-ray Images Based on Fusion Variational Auto-Encoder and Three-Terminal Attention Mechanism," Symmetry, vol. 13, no. 5, p. 814, 2021, doi:10.3390/sym13050814.

[26] Kim, M. y Lee, B.-D., "Automatic Lung Segmentation on Chest X-rays Using Self-Attention Deep Neural Network," Sensors, vol. 21, no. 2, p. 369, 2021, doi:10.3390/s21020369.

[27] Selvan, R., Dam, E. B., Detlefsen, N. S., Rischel, S., Sheng, K., Nielsen, M., y Pai, A., "Lung Segmentation from Chest X-rays using Variational Data Imputation," 2020, http://arxiv.org/abs/2005.10052 (visitado el 2022-12-17). arXiv:2005.10052 [cs, eess, stat].

[28] Dorri Giv, M., Haghighi Borujeini, M., Seifi Makrani, D., Dastranj, L., Yadollahi,

M., Semyari, S., Sadrnia, M., Ataei, G., y Riahi Madvar, H., "Lung Segmentation using Active Shape Model to Detect the Disease from Chest Radiography," Journal of Biomedical Physics & Engineering, vol. 11, no. 6, pp. 747–756, 2021, doi:10.31661/jbpe.v0i0.2105-1346.

[29] Wang, W., Feng, H., Bu, Q., Cui, L., Xie, Y., Zhang, A., Feng, J., Zhu, Z., y Chen, Z., "MDU-Net: A Convolutional Network for Clavicle and Rib Segmentation from a Chest Radiograph," Journal of Healthcare Engineering, vol. 2020, pp. 1–9, 2020, doi:10.1155/2020/2785464.

[30] Wessel, J., Heinrich, M. P., von Berg, J., Franz, A., y Saalbach, A., "Sequential Rib Labeling and Segmentation in Chest X-Ray using Mask R-CNN," 2019, http://arxiv.org/abs/1908.08329 (visitado el 2022-12-17). arXiv:1908.08329 [eess].

[31] Santosh, K., Candemir, S., Jaeger, S., Folio, L., Karargyris, A., Antani, S., y Thoma, G., "Rotation Detection in Chest Radiographs Based on Generalized Line Histogram of Rib-Orientations," en 2014 IEEE 27th International Symposium on Computer-Based Medical Systems, (New York, NY, USA), pp. 138–142, IEEE, 2014, doi:10.1109/CBMS.2014.56.

[32] Santosh, K. C., Candemir, S., Jaeger, S., Karargyris, A., Antani, S., Thoma, G. R., y Folio, L., "Automatically Detecting Rotation in Chest Radiographs Using Principal Rib-Orientation Measure for Quality Control," International Journal of Pattern Recognition and Artificial Intelligence, vol. 29, no. 02, p. 1557001, 2015, doi:10.1142/S0218001415570013.

[33] Dovganich, A. A., Khvostikov, A. V., Krylov, A. S., y Parolina, L. E., "Automatic Quality Control in Lung X-Ray Imaging with Deep Learning," Computational Mathematics and Modeling, vol. 32, no. 3, pp. 276–285, 2021, doi:10.1007/s10598-021-09539-6.

[34] Jaeger, S., Candemir, S., Antani, S., Wáng, Y.-X. J., Lu, P.-X., y Thoma, G., "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," Quantitative Imaging in Medicine and Surgery, vol. 4, no. 6, pp. 475–477, 2014, doi:10.3978/j.issn.2223-4292.2014.11.20.

[35] Nguyen, H. C., Le, T. T., Pham, H. H., y Nguyen, H. Q., "VinDr-RibCXR: A Benchmark Dataset for Automatic Segmentation and Labeling of Individual Ribs on Chest X-rays," 2021, http://arxiv.org/abs/2107.01327 (visitado el 2022-12-17). arXiv:2107.01327 [cs, eess].

[36] Bustos, A., Pertusa, A., Salinas, J.-M., y de la Iglesia-Vayá, M., "PadChest: A large chest x-ray image dataset with multi-label annotated reports," Medical Image Analysis, vol. 66, p. 101797, 2020, doi:10.1016/j.media.2020.101797. arXiv:1901.07441 [cs, eess].

[37] Alnaser, A., Gong, B., y Moeller, K., "Evaluation of open-source software for the lung segmentation," Current Directions in Biomedical Engineering, vol. 2, no. 1, pp. 515–518, 2016, doi:10.1515/cdbme-2016-0114.

[38] Cohen, J. P., Viviano, J. D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M. P., Chaudhari, A., Brooks, R., Hashir, M., y Bertrand, H., "TorchXRayVision: A library of chest X-ray datasets and models," 2021, http://arxiv.org/abs/2111.00595 (visitado el 2023-02-26). arXiv:2111.00595 [cs, eess].

ANNEXES

A. Additional Results

A.1. Visualization of Cases

A.1.1. Image Sample

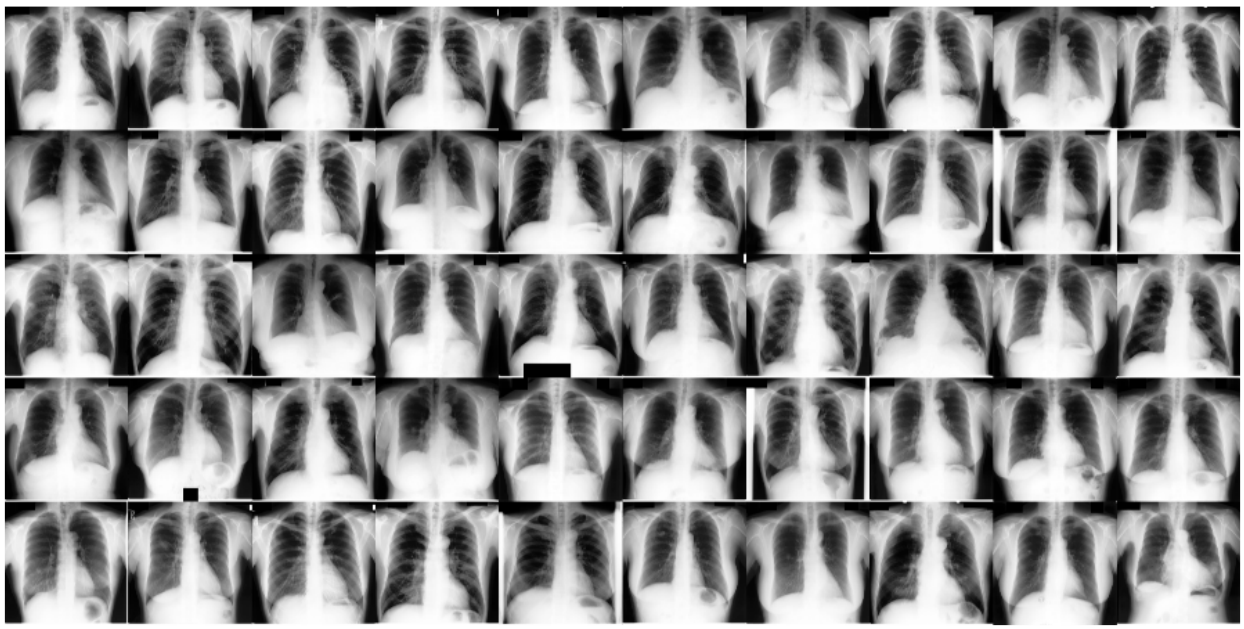


Figure A.1: Evaluation Image Subset

A.1.2. Complex Cases in Lung Segmentation

Complex cases in lung segmentation usually correspond to tuberculosis presenting cases. In figure A.2 we sample a group of complex cases. As you can determine, lowermost borders remain untouched, not influencing on the medial line determination algorithm.

Figure A.2: Complex Cases in Lung Segmentation

A.1.3. Heart Segmentation Sample

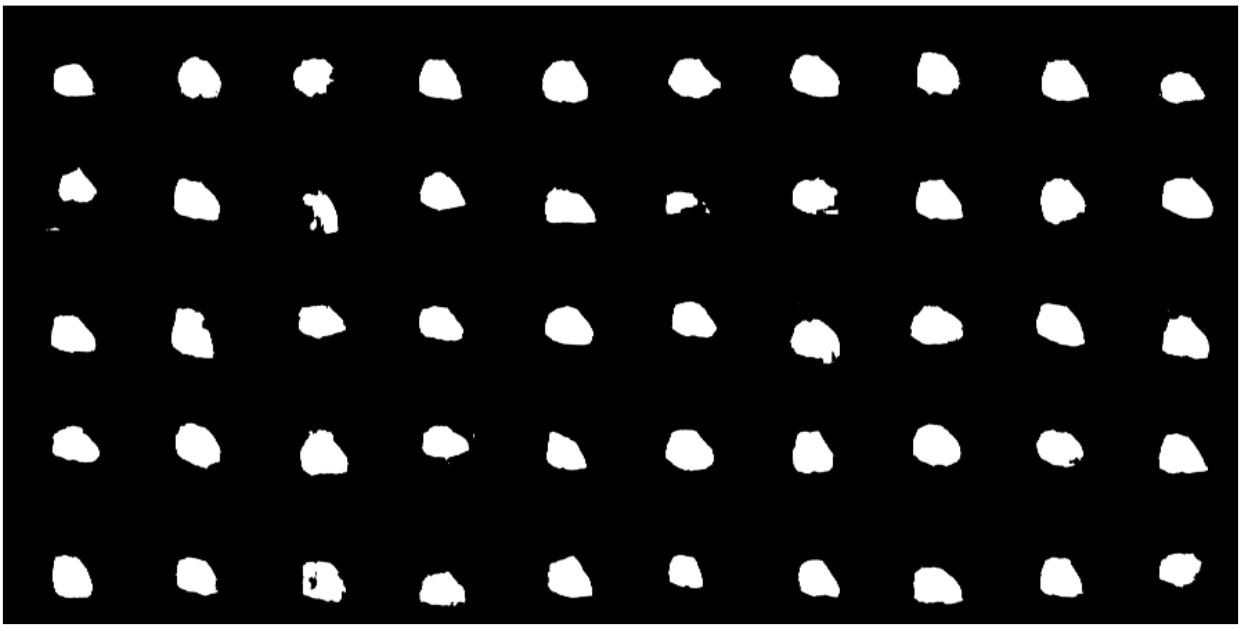


Figure A.3: Heart Segmentation Sample

A.1.4. Rib Segmentation Sample

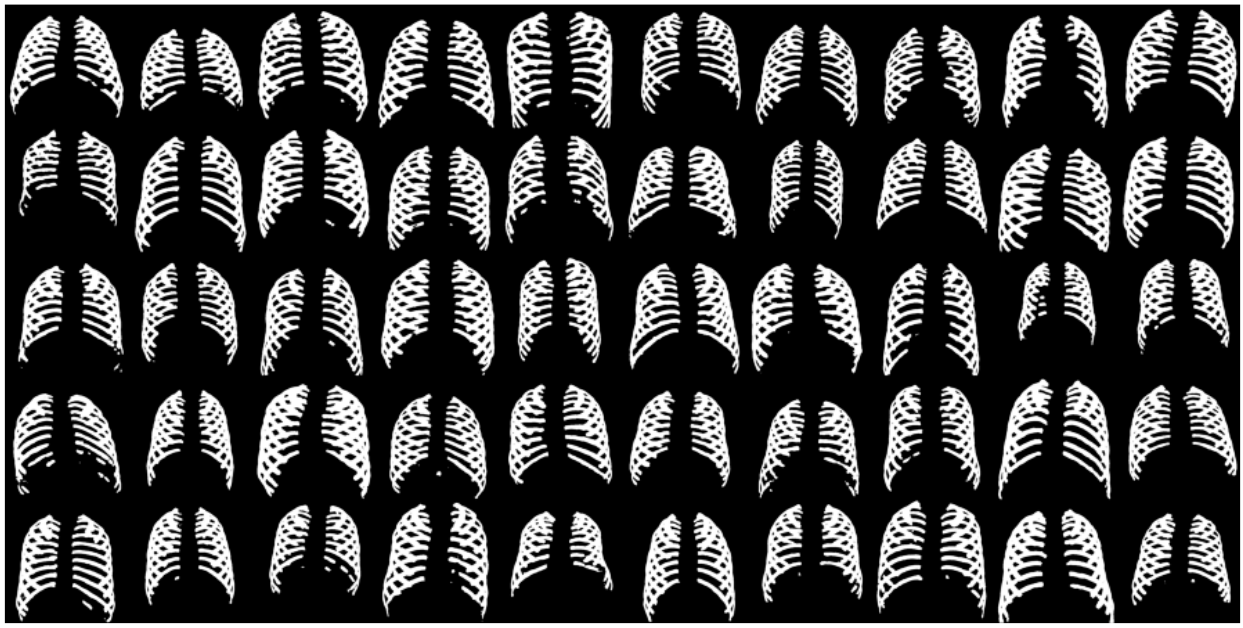


Figure A.4: Rib Segmentation Sample

A.1.5. Clavicle Segmentation Sample

Figure A.5: Clavicle Segmentation Sample

B. Architecture Representations

B.1. UNet

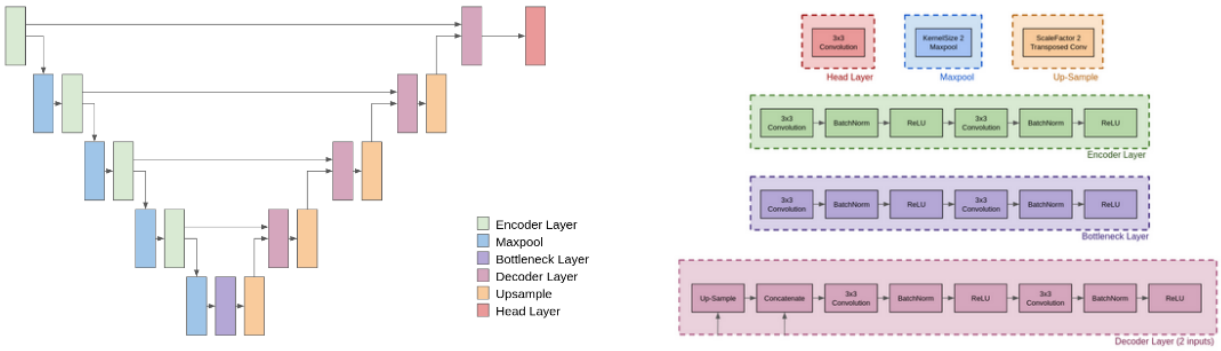


Figure B.1: U-Net Model

B.2. ResNet-UNet

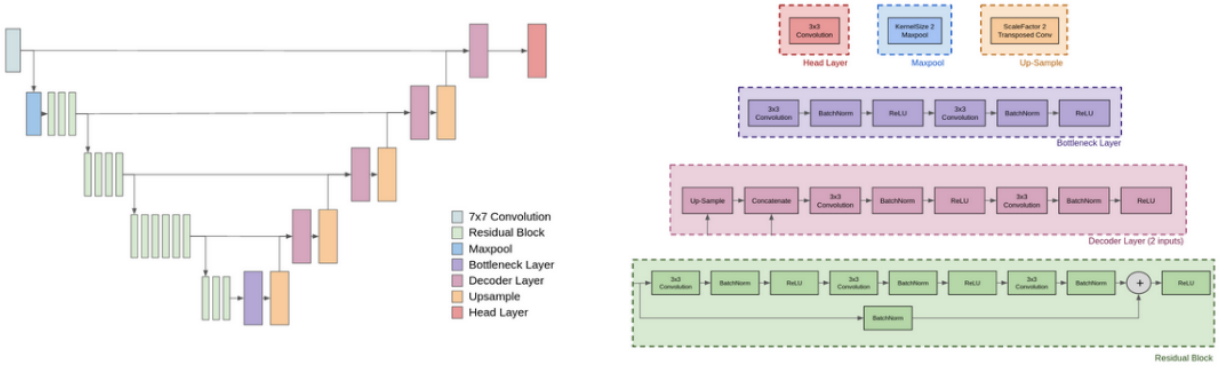


Figure B.2: ResNet-UNet Model

B.3. Swin-UNet

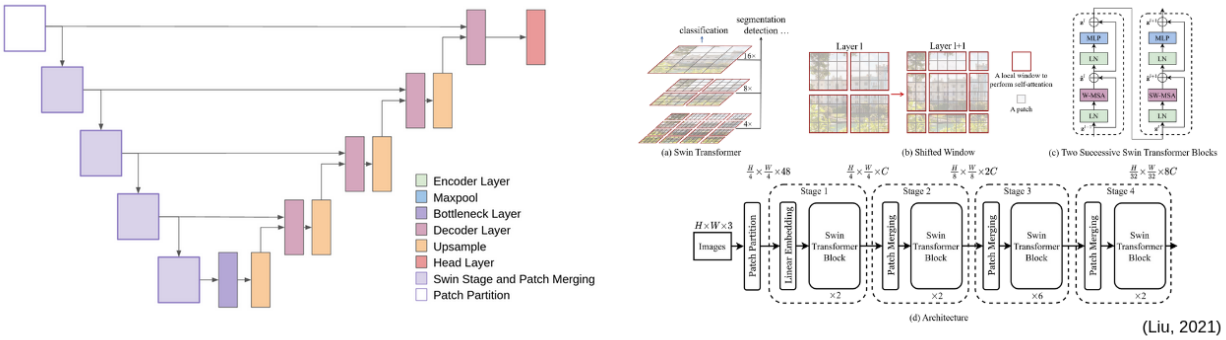


Figure B.3: Swin-UNet Model

B.4. Three Head Attention U-Net

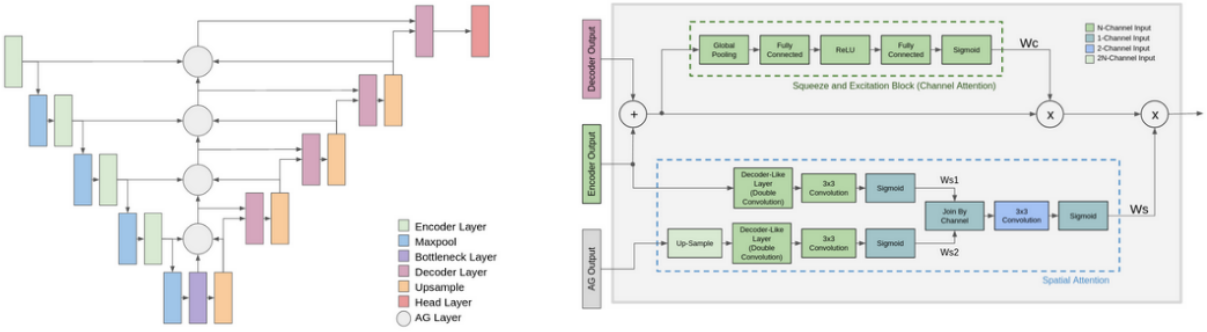


Figure B.4: Three Head Attention U-Net Model

B.5. Spatial Attention U-Net

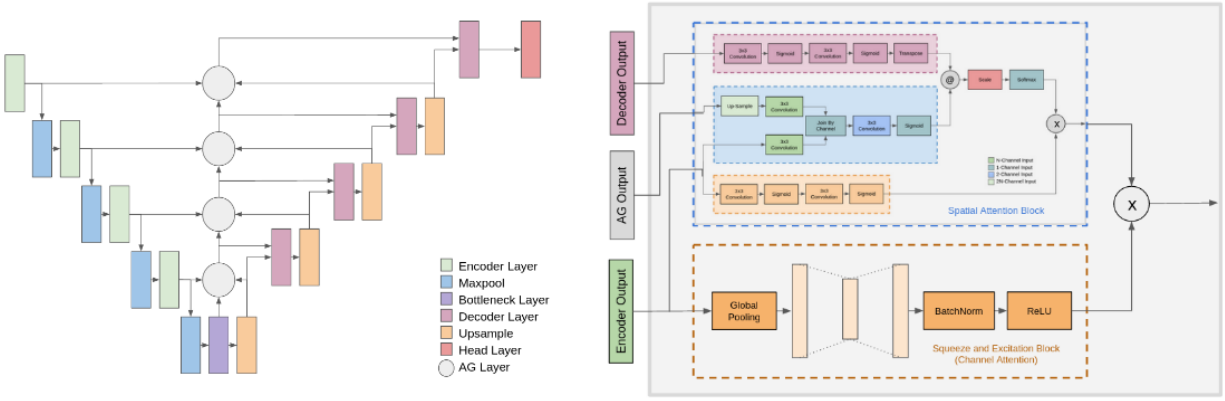


Figure B.5: Spatial Attention U-Net Model

B.6. Double Spatial Attention U-Net

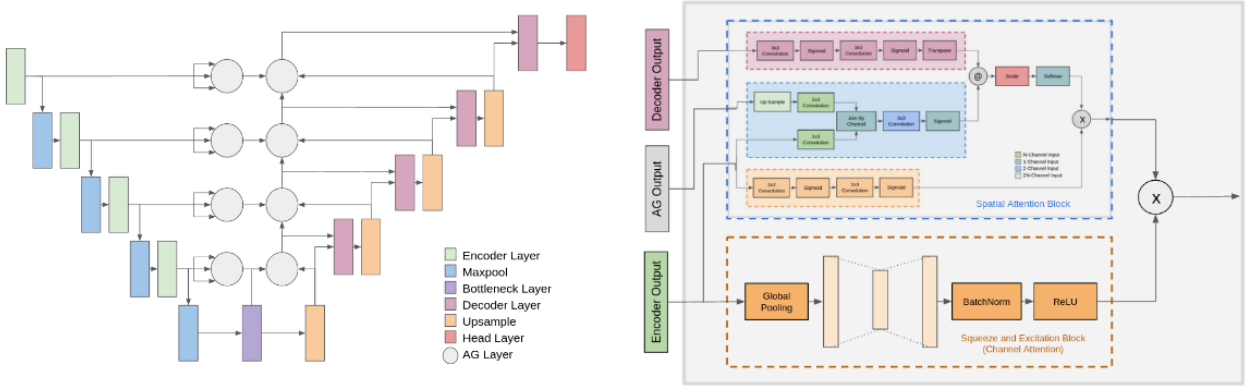


Figure B.6: Double Spatial Attention U-Net Model

B.7. Full Spatial Attention U-Net

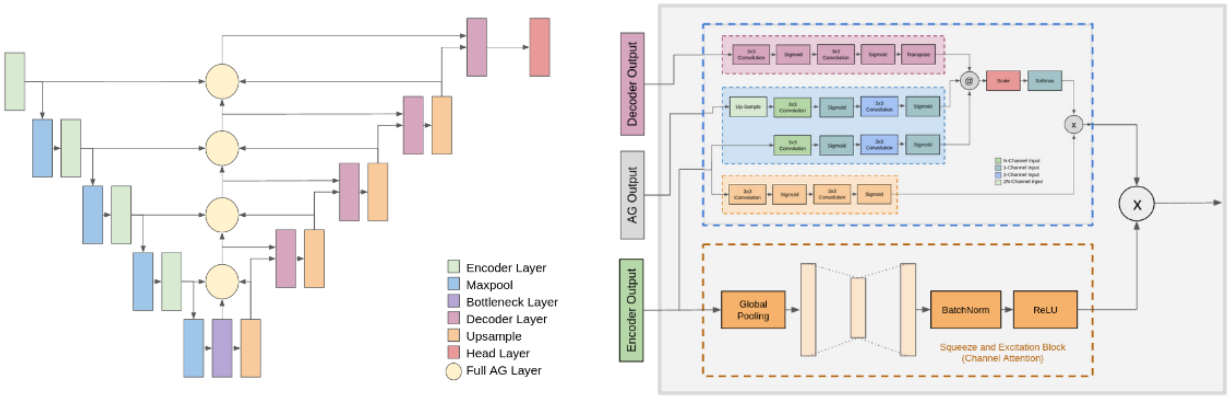


Figure B.7: Full Spatial Attention U-Net Model

B.8. Spatial Decoder U-Net

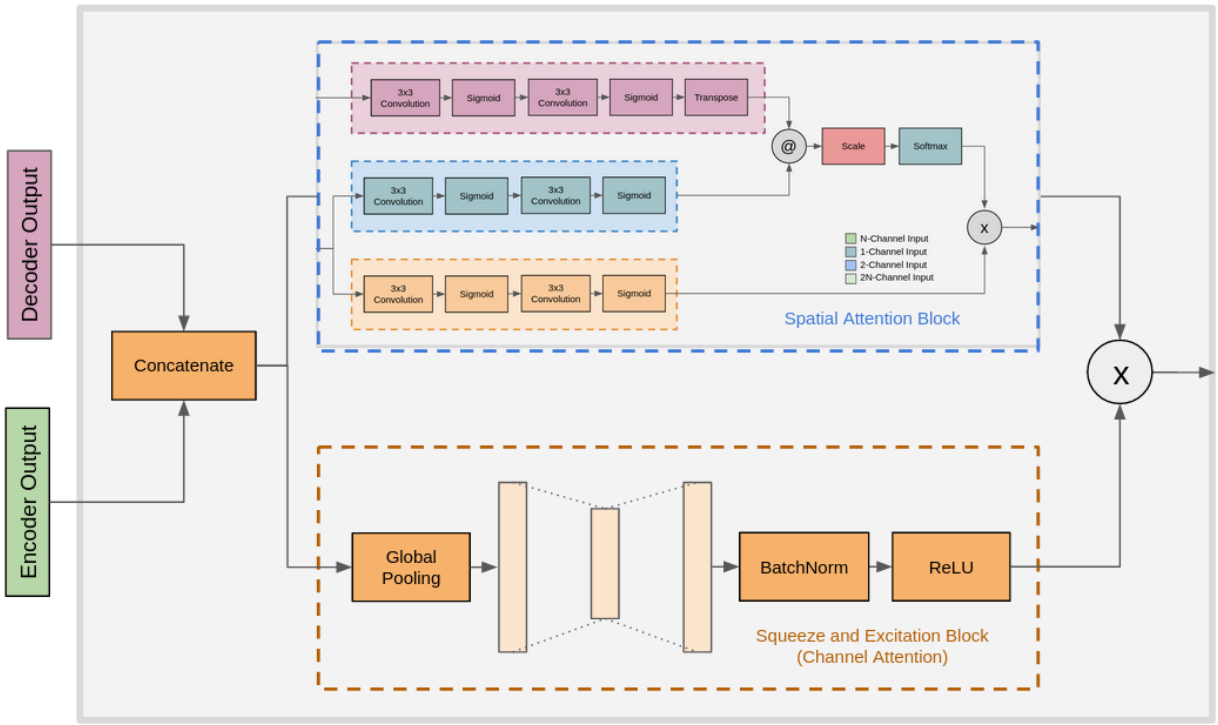


Figure B.8: Spatial Decoder U-Net Attention Block