MULTI-TASK NETWORK FOR COMPUTED TOMOGRAPHY SEGMENTATION THROUGH FRACTAL DIMENSION ESTIMATION

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER ENGINEERING

By

Aziza Saber Jabdaragh

January 2023

Multi-Task Network for Computed Tomography Segmentation Through Fractal Dimension Estimation
By Aziza Saber Jabdaragh
January 2023

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Çiğdem Gündüz Demir(Advisor)

Selim Aksoy(Co-Advisor)

Shervin Rahimzadeh Arashloo

Tunca Doğan

Approved for the Graduate School of Engineering and Science:

Orhan Arıkan
Director of the Graduate School

ABSTRACT

MULTI-TASK NETWORK FOR COMPUTED TOMOGRAPHY SEGMENTATION THROUGH FRACTAL DIMENSION ESTIMATION

Aziza Saber Jabdaragh
M.S. in Computer Engineering
Advisor: Çiğdem Gündüz Demir
January 2023

Multi-task learning proved to be an effective strategy to increase the performance of a dense prediction network on a segmentation task, by defining auxiliary tasks to reflect different aspects of the problem and concurrently learning them with the main task of segmentation. Up to now, previous studies defined their auxiliary tasks in the Euclidean space. However, for some segmentation tasks, the complexity and high variation in the texture of a region of interest may not follow the smoothness constraint in the Euclidean geometry. This thesis addresses this issue by introducing a new multi-task network, *MTFD-Net*, which utilizes the fractal geometry to quantify texture complexity through self-similar patterns in an image. To this end, we propose to transform an image into a map of fractal dimensions and define its learning as an auxiliary task, which will provide auxiliary supervision to the main segmentation task, towards betterment of left atrium segmentation in computed tomography images. To the best of our knowledge, this is the first proposal of a dense prediction network that employs the fractal geometry to define an auxiliary task and learns it in parallel to the segmentation task in a multi-task learning framework. Our experiments revealed that the proposed *MTFD-Net* model led to more accurate left atrium segmentation, compared to its counterparts.

Keywords: Fractal dimension, multi-task learning, dense prediction networks, fully convolutional networks, computed tomography, segmentation, left atrium.

ÖZET

FRAKTAL BOYUT TAHMİNİ KULLANARAK BİLGİSAYARLI TOMOGRAFİ SEGMENTASYONU İÇİN ÇOK-GÖREVLİ AĞ

Aziza Saber Jabdaragh
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Çiğdem Gündüz Demir
Ocak 2023

Çok-görevli öğrenmenin, problemin farklı yönlerini yansıtmak üzere yardımcı görevleri tanımlamasının ve bunları ana görev olan segmentasyon ile eş zamanlı olarak öğrenmesinin, yoğun bir tahmin ağının segmentasyon görevindeki performansını artırmak için etkili bir strateji olduğu kanıtlanmıştır. Önceki çalışmalar, yardımcı görevlerini Öklid uzayında tanımlamışlardır. Ancak, bazı bölütleme görevlerinde, ilgi bölgesinin dokusundaki karmaşıklık ve yüksek değişkenlik, Öklid geometrisindeki düzgünlük kısıtlamasına uymayabilir. Bu tez, bir görüntüdeki doku karmaşıklığını birbirine benzer desenler aracılığıyla nicelleştirmek için fraktal geometriyi kullanan yeni bir çok-görevli ağ olan *MTFD-Net* modelini tanımlayarak bu sorunu ele almaktadır. Bu amaçla, bir görüntüyü fraktal boyut bir haritasına dönüştürmeyi ve bunun öğrenilmesini yardımcı bir görev olarak tanımlamayı öneriyoruz. Bu, bilgisayarlı tomografi görüntülerinde sol kulakçık segmentasyonunun iyileştirilmesine yönelik ana segmentasyon görevine yardımcı denetim sağlamak amacıyladır. Bildiğimiz kadarıyla, bu, bir yardımcı görevi tanımlamak için fraktal geometriyi kullanan ve bunu çok-görevli bir öğrenme modelinde, bölütleme görevine paralel olarak öğrenen yoğun tahmin ağının ağı öneren ilk çalışmadır. Deneylerimiz, önerilen *MTFD-Net* modelinin, benzerlerine kıyasla daha doğru sol kulakçık segmentasyonuna olanak sağladığını ortaya koymuştur.

Anahtar sözcükler: Fraktal boyut, çok-görevli öğrenme, yoğun tahmin ağları, tam evrişimli ağlar, bilgisayarlı tomografi, segmentasyon, sol kulakçık.

Acknowledgement

I would like to express my deepest gratitude and special thanks to my supervisor, Prof. Dr. Ciğdem Gündüz Demir, without whom I would not have been able to complete this research or earn my master's degree. Her generous support, invaluable assistance, heartfelt kindness, unwavering patience, and unfailing encouragement always created immense enthusiasm in me to undertake my master's study and successfully write my thesis. She has been the ideal advisor in every way.

Additionally, this endeavor was only possible thanks to the kind examination and constructive criticism of my defense committee, Prof. Dr. Shervin Rahimzadeh Arashloo and Prof. Dr. Tunca Doğan, and my co-supervisor, Prof. Dr. Selim Aksoy, who reviewed my thesis and provided feedback. I am expressing my sincere gratitude to Bilkent University, librarians, research assistants, professors, and academic and non-academic staff, who generously provided an appropriate atmosphere, knowledge, and expertise. I thank the study participants who provided the necessary data and crucial information. I am also thankful for the partly support of this thesis by the Scientific and Technological Research Council of Turkey through the project TÜBİTAK 220N354.

Lastly, words cannot express my appreciation to my parents, Sarfnaz Sadr and Ali Saber Jabdaragh, and my siblings, who have always believed in me in failure and success. I have learned hard work from my parents, and their belief and confidence in me have kept my spirits and motivation high. Completing my study journey would have been impossible without their continuous and priceless support, understanding, and encouragement. Additionally, I want to express my gratitude and sincerity to my friends who made life and study enjoyable, especially at Bilkent University. They filled the void left by my family's and my country's absence with their affection and emotional support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Atrial fibrillation is a disease that affects the left atrium (LA) of the heart. Its manual segmentation [1, 2] requires highly experienced labor effort while being an error-prone and time-consuming process, which becomes challenging when there are large volumes of patients. Thus, its automated segmentation becomes an important tool. To this end, deep learning methods have been proposed. These methods outperform traditional machine learning methods, typically yielding higher performance in this automated segmentation task. The atrial segmentation challenge in 2018 [3] confirms this fact since most of the methods in this challenge that used deep learning methods achieved statistically significantly higher accuracies compared to traditional multi-atlas-based methods. However, LA is known for its variation in shape, size, and anatomical complexity among patients and it is unique for different individuals. Thus, it still needs to be addressed for existing deep learning approaches to segment LA accurately.

This thesis aims to describe a novel technique for LA segmentation in computed tomography (CT) scans by introducing a multi-task network that we named *MTFD-Net*. It is different than the previous networks since they calculated their descriptors in the Euclidean space, which does not fully represent the complexity and high variation in the texture of a region of interest (ROI) needed in CT images. We propose to define a richer set of descriptors that deal with structures

that are not exactly Euclidean [4] while better representing irregular textures and rugged surfaces using fractals. Our network concurrently learns an auxiliary regression task to estimate fractal maps generated on each image pixel-wise and the main segmentation task to estimate LA segmentation maps. Maps of fractal dimensions (FD) can model complex geometric patterns of texture through self-similarity attribute in CT images. Our model was evaluated on an LA dataset and achieved higher performance than single-task and shape-prior multi-task networks.

1.1 Motivation

Image segmentation is essential for computer-aided diagnosis systems in different applications. Despite the considerable research work in this area, it remains a challenging problem due to high variability in the morphological appearance of the images and imaging modalities such as CT, X-ray, and magnetic resonance imaging (MRI). CT imaging is one of the most widely used radiographic techniques in diagnosis, clinical studies, and treatment planning due to its high spatial resolution and ease of use. Accurate heart segmentation of patients is one of the critical segmentation tasks to detect cardiovascular-related illnesses such as atrial fibrillation. The left atrium is one of the four chambers in the heart with a very complex anatomical structure, which transfers oxygenated blood from the lungs to the left ventricle. An abnormality causes atrial fibrillation in this pumping process, which causes a very irregular and relatively faster pulse rate in the heart, causing the LA to quiver. Atrial fibrillation can cause blood clotting in the LA, leading to stroke. Therefore, LA segmentation is necessary for evaluating atrial size and function, which are impertinent for treating diseases such as atrial fibrillation.

Encoder-decoder networks have shown great promise for many medical image segmentation problems, including the segmentation of CT images. Nevertheless, it is still challenging for these networks to segment an ROI with high texture variations and complexity, especially when the complexity and irregularity of the

texture make the boundaries between the ROI and its surrounding tissues partially or entirely indiscernible. To alleviate this problem and, thus, to better segment the ROIs, many studies proposed to employ contextual, shape, and contour information in their network design. One group of these studies used the attention mechanism to preserve shape and spatial information [5, 6]. Another group exploited this information in the form of designing a multi-task network. They typically defined learning contour [7] and shape [8, 9] related descriptors as auxiliary tasks and trained their networks to concurrently learn these tasks with the main task of ROI segmentation. Despite their success in various applications, all these studies calculated their descriptors in the Euclidean space. However, in a CT image, the complexity and high variation in the texture of an ROI, especially when it is a part of a soft organ and contains intralesional heterogeneity, may not follow the smoothness constraint in the Euclidean geometry. On the contrary, it is possible to define a richer descriptor set for irregular textures and rugged surfaces using fractals, which deal with structures that are not exactly Euclidean [4].

Measuring the interaction and correlation of pixels with their surrounding pixel intensities may provide useful information about the whole image since it is shown that this spatial dependence of the gray-level texture gives valuable information about spatial distribution [10]. Such texture properties can be used to distinguish tissues in a CT image from each other that are hard to discriminate using previous methods. The fractal dimension is a good representative of texture composition to adequately preserve the spatial information of each gray-level intensity with respect to its neighboring pixels through a quantitative measurement and, thus, models the complexity of texture. Moreover, variation in ROIs may be addressed more effectively since fractal geometry treats objects of different sizes and shapes similarly when they follow the same self-similar patterns at different scales [11]. Many studies have shown the successfulness of fractal geometry in many applications, such as classification and segmentation. Such studies show that using fractal dimension alone does not provide sufficient information for an accurate description of natural textures while combined with other kinds of features resulted

in higher performance [12, 13]. However, none of those studies used fractal geometry in a deep learning model. Thus, using FD richer descriptors may improve the performance for segmentation problems in a deep learning model.

1.2 Contributions

This thesis presents a new dense prediction neural network design, which we name *MTFD-Net*. In this design, we propose calculating fractal dimension maps to quantify texture complexity through self-similar patterns in a CT image and introduce an end-to-end framework that utilizes these maps through a multi-task network for LA segmentation in CT images. This approach is applicable to different deep learning architectures since the fractal dimension information is used in the output map of the proposed network. The main contributions of this thesis are two-fold:

- *MTFD-Net* transforms a CT image into a map of fractal dimensions to model the texture complexity of an LA region. This transformation results in a representation complementary to the ground truth map, making it easier to segment the LA regions [14]. It then defines the learning of this fractal dimension (FD) map as an auxiliary task in the network design to provide auxiliary supervision to the main task of LA segmentation. Although there exist studies that used fractals solely or together with other handcrafted features in the design of a traditional classification and segmentation model [14, 15, 16], these studies did not utilize fractal textures in a dense prediction neural network design.

- *MTFD-Net* proposes to learn the auxiliary task of FD map estimation together with the main task of LA segmentation in a multi-task learning framework. To this end, it constructs a network with a shared encoder path and two decoder paths, one for each task, and end-to-end learns these two tasks simultaneously. This multi-task learning framework has two main

benefits. First, concurrent learning of the two tasks from a single shared encoder requires learning a shared representation that works adequately well for both of these tasks. This is known as an effective means to decrease the likelihood of each task overfitting and hence, to obtain more generalized models since it is more difficult to finetune one representation on two different tasks at the same time [17]. Second, the shared feature representation learned at the various layers of the shared encoder path should keep the necessary context and texture information to realize LA segmentation and FD map estimation, respectively. This forces the network to effectively incorporate additional texture information into the segmentation process. Although there exist studies that developed multi-task network architectures for various applications [7, 8, 9, 18, 19], none of them defined their auxiliary task using the fractal geometry.

1.3 Outline

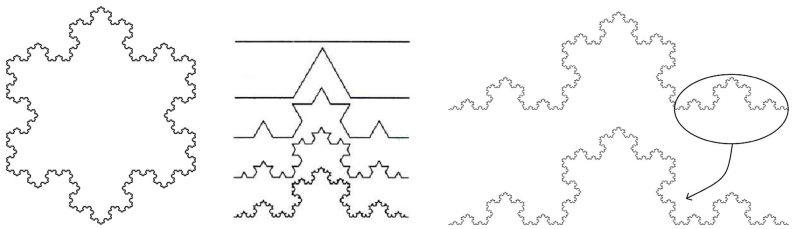
The outline of this thesis is as follows. In Chapter 2, fractal dimensions are reviewed, and utilization of fractal dimensions in medical image analysis and different approaches for computing the fractal dimension of medical images are briefly explained. Chapter 3 introduces the details of our proposed approach. Chapter 4 provides all the information regarding our experimental dataset, implementation details, evaluation metrics, post-processing, and comparison methods. It also provides the quantitative and visual results for our model and the comparison methods. Chapter 5 concludes this thesis and discusses its potential future directions.

Chapter 2

Background

The fractal dimension was first introduced by [4], motivated by Hausdorff's work from 1919, to describe structures that show similar patterns at different scales which cannot be described in the Euclidean geometry. Since then, it has been widely used in many applications, including medical image segmentation [15]. This characteristic is known as self-similarity, the complexity of which is typically quantified by the fractal dimension. A structure in a space is called self-similar if each small piece of it has the same pattern. If such small pieces were copied, the overall shape could be obtained, and vice versa the small pieces could be obtained by applying a self-similarity transformation to the overall structure.

The whole idea of the FD is that the measured length, area, or volume of objects and surfaces follows a function of changing their scale. Mandelbrot described an approach in 1967 to measure the length of the Coast of Britain while using different scales of measurement and estimating the FD [20]. He estimated the total length of a crooked coastline at a given spatial scale as a set of straight lines with the same length. He realized that when increasing the spatial scale, the measured length increases while the details of the coastline increase or decrease. Thus, it could be concluded that the areas, surfaces, lengths, and volumes could be so complex that the standard measurements in the Euclidean distance space would be meaningless.



(a) (b) (c)

Figure 2.1: An example of a fractal structure and self-similarity. (a) Koch snowflake. Figure taken from `https://personal.math.ubc.ca/~cass/`. (b) Koch curve and steps of its construction. Figure taken from [24]. (c) Koch curve self-similarity property. Figure taken from [25].

Mandelbrot discovered a method for measuring the degree of complexity while taking into account the pace of growth in the surface, length, or volume with respect to ever-smaller scales. Such measurements are explained by a quantity D that includes the attributes of a dimension in a fractional manner that cannot be described in standard Euclidean dimensions. More precisely, such surfaces had a dimension between Euclidean dimensions (non-integer) that could not be characterized by the Euclidean distance [21, 22, 23]. To this end, he assumed that rather than altering arbitrarily with scale changes, there is a function for quantifying the relationship between length, surface, or volume with their scale. Such a function will lead to finding more correct dimensions. Such curves, lengths, and areas are also considered to be statistically self-similar, as is the case in many natural objects and geographical regions that usually follow similar behaviors at different scales. Statistically self-similarity refers to the fact that each component of the object, when scale taken into account, would follow a pattern that is either identical to or very similar to the whole object. One well-known fractal shape is the Koch snowflake shown in Figure 2.1.

The fractal dimension, in which a shape is no longer a two-dimensional structure, could be used to measure the overall structure's detail. Instead, it is defined by an infinite number of lines representing a one-dimensional space. At the same time, it expands the structure into a shape characterized by two-dimensional space. Thus, the dimension of the overall structure (representative of the complexity) is between 1D and 2D. To calculate this dimension, consider the scaling

factor r for generating the Koch curve in Figure 2.1. It can be seen that each line is divided into three parts, creating four copies of the original structure, while an equilateral triangle is added in the middle at each repetition. Thus, the number of copies, $N = 4$, can be determined by 3 to the power of dimension D where the scaling factor $r = 1/3$. Therefore, dimension D for this curve can be obtained as in Equation 2.1.

$$4 = 3^D \quad \Rightarrow \quad D = \frac{\log 4}{\log 3} = 1.26 \tag{2.1}$$

This shows what we claimed about the structure's dimensionality. In general, the FD of a structure could be obtained by Equation 2.2.

$$N = (1/r)^{FD} \quad \Rightarrow \quad FD = \frac{\log N}{\log 1/r} \tag{2.2}$$

The remaining parts of this chapter are divided into three sections: 1) Fractal analysis in medical imaging, which mentions and briefly discusses earlier work on applying the FD in the medical image domains, particularly in texture-related segmentation problems. 2) Fractal dimension computing methods, which explains some well-known FD estimation techniques and their improvements and variations. 3) Network architectures, which provides a description of the cutting-edge deep models that have been applied to various segmentation problems for medical image analysis.

2.1 Fractal Analysis in Medical Imaging

Fractal dimensions, in particular, have been widely used in pattern recognition and medical image analysis challenges, particularly in texture segmentation and classification. Fractal geometry is a texture analysis technique. In one of the earliest efforts [16], the differential box-counting approach was introduced and

utilized to segment textures using fractal geometry. In [26], the authors proposed a method to test the vast range of variations in the fractal dimension of internal and peripheral textures of small peripheral bronchogenic carcinomas in thin-section computed tomography. They discovered that the obtained FD for the internal texture and peripheral textures were substantially different for two cases after dividing bronchogenic carcinomas into bronchioloalveolar cell carcinomas (BACs) and non-BACs.

The fractal features were also applied for liver tissue classification problems along with M-band wavelet transform [27]. Similarly, the FD was utilized in another study for liver classification [28]. The work in [29] used the fractal dimensions in both geometric and stochastic fractal for abnormality classification in human lung CT scans. The authors of [30] used FDs to classify breast ultrasound images into benign and malignant, and they demonstrated promising results by applying the fractal geometry on enhanced images. In another attempt, statistical fractal-dimension features were used to enhance the discrimination between benign and malignant breast masses [31].

The majority of natural objects and surfaces exhibit statistical self-similarity or self-affine properties that can be examined by fractal theory, according to these studies and additional research. This is predicated on the notion that textures with finite resolutions have self-similarity properties. Such characterization by the fractal geometry is called to be scale-invariant, which reflects the complexity and irregularity of the objects [32]. A finite resolution image's fractal dimension is expressed as $\frac{\log N_r}{\log 1/r}$, where N_r is the number of boxes required to overlay the entire image and r is the constant applied to different scaling factors. Later explanations and discussions will cover various methodologies and mathematical derivations.

2.1.1 Fractal Dimension Computing Methods

The degree of complexity can be measured by the increment speed of a length, surface, curve, and volume when the shape's scale decreases by a factor measured by the fractal dimension. When a structure is self-similar, it is easier to compute the FD, as the overall structure is a compass of patterns. Though, in more complex and ambiguous structures, it is challenging to measure the dimensionality of the structure. To this end, there are different techniques, including box-counting, variance, and spectral methods, described in [33], out of which box-counting has given the most successful results in transforming the intensities of images into fractal dimensions. Thus, its improved versions and variants have been introduced [34, 35, 36, 37]. The box-counting method, equally applicable to objects and forms in space, aims to quantify the FD in a plane with any structural pattern, with or without self-similar patterns. Since an exact answer for the Hausdorff-Besicovitch equation (proportional equation for calculating FDs) cannot be calculated, different methods result in different FD values for the same features or texture. Thus, all methods try to estimate it by following three steps: first, they measure the quantities of the objects in different scales like in Britain's Coast measurement. Second, they plot the log of measured quantity versus the log of scale and fit a regression line [38] to the obtained data points. Finally, the FD is estimated by the slope of the regression line as $FD = \frac{\log N_r}{\log 1/r}$ where N_r is the number of smaller subsets of the overall self-similar structure of the image and r is the corresponding scaling factor, that is, the factor needed to observe self-similar parts.

In this work, to capture the roughness of arbitrarily sized images, we have used the differential box-counting method (DBC) [39], an adapted version of the box-counting method, for computing FDs of the images as it has become a choice of research in FD computation. This method could be applied to gray-scale images [16], and it is more efficient than the original box-counting method and its variations. DBC considers the minimum and maximum pixel intensity values of the images [13, 36, 37, 40, 41]. In recent literature, different versions of both box-counting and DBC approaches were used to estimate the FD of both

gray-scale and color images [22, 42, 43, 44]. This thesis uses the differential box-counting method to calculate the fractal dimension maps for gray-scale images. The details will be given in the next chapter.

2.2 Deep Learning Models

Fully convolutional networks (FCN) were first introduced in [45] for an end-to-end, pixel-wise image segmentation task network using encoder-decoder paths. VGG-16 [46] without fully connected layers at the end and a decoder module for full dense prediction were appended for the encoder part. The idea for the decoder part was that the coarse features extracted in the encoder part have minimal spatial resolution due to pooling layers added to it. Thus, they need to be upsampled and refined in the decoder part for final dense prediction. Transpose convolutional layers were used for upsampling down-sampled feature maps into a full-resolution segmentation map. However, designing a network only by adding convolution and pooling layers in the encoder part and upsampling layers in the decoder part does not result in fine-grained segmentation maps. This is because, despite sophisticated techniques like transpose convolution, adding max-pooling in shallower layers causes the spatial position information to be lost at the expense of gaining precise semantic contextual information. Therefore, a new design is needed to combine these two pieces of information for better segmentation. To this end, skip connections are added to the network to slowly upsample features of the encoder part while fusing them with locational features in deeper layers of the decoder module to cover finer details of shape and boundarie.

The *U-Net* architecture [47] is one of the well-known primary FCNs that expanded the capacity of decoder modules by adding skip connections corresponding to each convolution block in the encoder part to the decoder part. The task was to segment cell and background as a two-class classification problem in gray-scale microscopy images of cells. *U-Net* consists of a contracting path (encoder) to capture context and an expansive path (decoder) to capture localization. The

network takes an input gray-scale microscopy image and produces a binary segmentation map (1 cell, 0 background).

Figure 2.2 shows the overall structure of *U-Net*. In the encoder module, a low-dimensional image representation is obtained using a traditional convolutional neural network like *VGG/ResNet*. This half of the architecture consists of blocks of convolution where each block is obtained by two convolutions followed by a rectified linear unit (*ReLu*) and then by a max-pooling layer for downsampling. The intermediate or central part of the figure represents the bottleneck structure. This part consists of feature maps with the lowest dimension, which basically are convolution layers without changing spatial size.

In the decoder module, discriminative or low-resolution features learned in the encoder module are semantically mapped onto high-resolution pixel space for final dense prediction. This module is the second half of the figure, which consists of upsampling using transposed convolutions, and concatenation followed by regular convolutional layers. There are skip connections at every pooling (mentioned as gray lines in the figure) stage where high-resolution feature maps from the previous pooling step of the encoder path are fused with upsampled feature maps in the decoder path through concatenation operations. The process is repeated until the original image spatial dimension is achieved. The primary purpose of skip connections is to feed one layer's output to another while skipping a few layers in between. 1×1 convolution layers are applied to map 64-channel feature vectors into the desired number of classification classes. The network outputs a probability map converted to a final prediction map through the softmax operation.

The *U-Net* architecture [47] was one of the first successful encoder-decoder networks and the most commonly used one for medical image segmentation. Although it achieved promising results for various segmentation problems, it might fail to accurately segment ROIs with texture variations and precisely delineate irregular ROI boundaries. Thus, many studies proposed *U-Net* variants that employed contextual, shape, and contour information to estimate better ROIs.

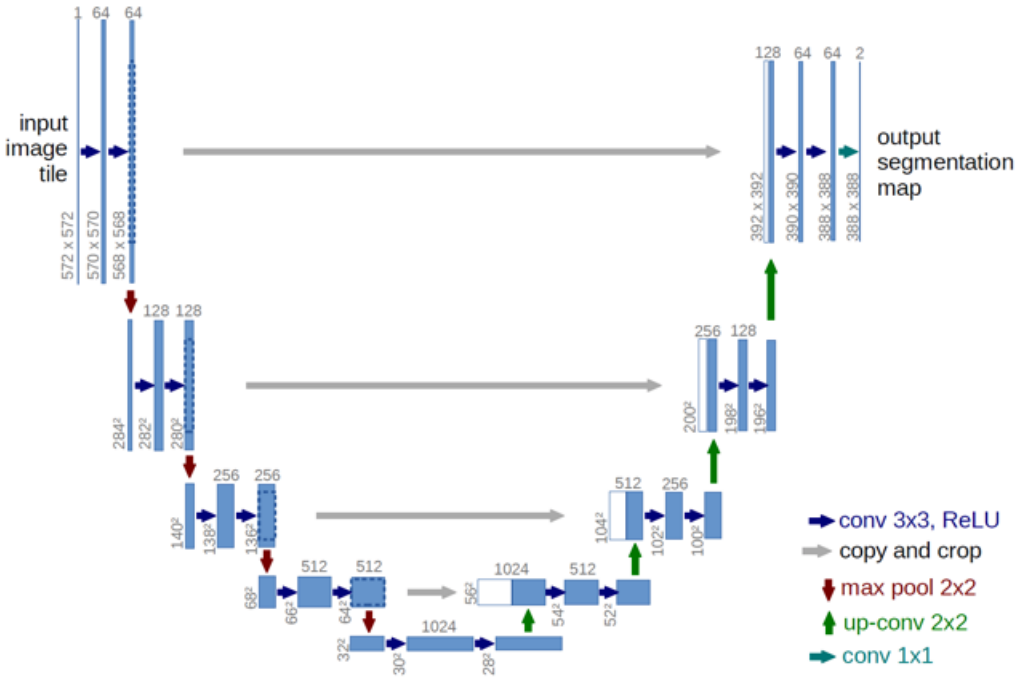


Figure 2.2: *U-Net* architecture illustration, with an encoder path (left side) that reduces the spatial resolution of the input slices and a decoder path (right side) that restores the original dimension. The blue and white boxes indicate multi-channel feature maps and concatenated copied feature maps from the encoder path (dashed boxes), respectively. The arrows are representative of different operations stated in the right legend. The numbers on the lower corner and top of each box denote image dimension (height and width) and the number of channels at each stage, respectively. Figure taken from [47].

One group of these studies used the attention mechanism, introduced into semantic segmentation by [48], to preserve shape and spatial information and used it in medical images [49] where they incorporated soft attention idea [50] in a *U-Net* backbone for a segmentation problem. A more state-of-the-art model, attention *U-Net* [51], was proposed in which attention gates were added to a standard *U-Net* backbone to preserve the spatial contextual information for pancreas segmentation in CT images. In this approach, the encoder path was unchanged, while attention gates were added in the decoder module into skip connections of *U-Net* to filter the salient features propagated through the skip connections. The idea was that some parts of input images were more important and relevant to

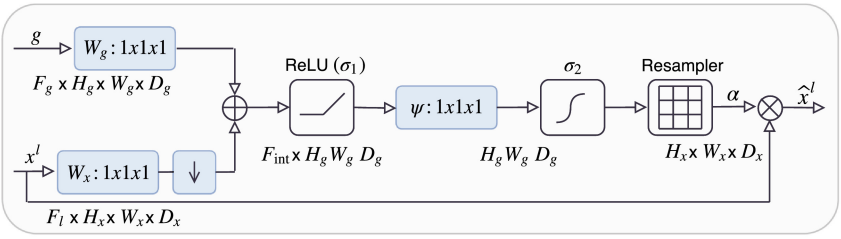


Figure 2.3: Attention gate (AG) schematic. Figure taken from [49].

ROI segmentation than others. Thus, instead of giving the same learning weight to all areas, it is reasonable for a network to pay more attention (give more weight) to the areas that incorporate more, or has high relevance, for the betterment of segmentation than other areas. The attention mechanism has different forms and is embedded in a network differently depending on the application. In the case of medical image segmentation in networks like [49], additive soft attention gates were added to the network, described in the next paragraph. In soft attention, the areas of the image with high relevance were multiplied with larger weights compared to low relevance areas. Thus, the model focused more on learning high-relevance areas as training proceeded.

Figure 2.3 represents an attention gate mechanism used in attention *U-Net*. Let g represent the input global feature vector that contains contextual information fed to the decoder part from the encoder part, and x represent the skip connection input coming from the decoder path. First, g and x are fed into 1×1 convolutions and upsampled. Then, they are fused through summation and passed through an element-wise nonlinearity *ReLu* activation layer. Finally, another 1×1 convolution is applied to it and normalized through a normalization function sigmoid to limit values into the $[0, 1]$ range. The output attention map is multiplied by the skip connection input for the final output of the attention module to focus on targeted areas while suppressing feature activation in unrelated areas.

These gates are added to skip connections since features extracted from the encoder module are normally directly passed through skip connections and concatenated with the features from the decoder module. Nevertheless, not all such

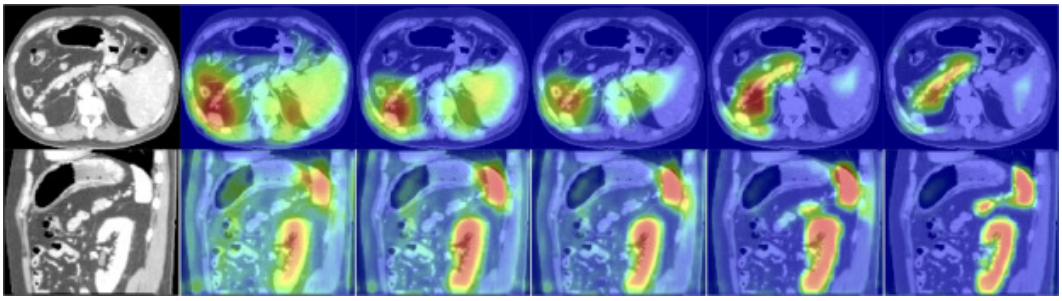


Figure 2.4: Attention coefficients (red highlights a higher attention coefficient) during different training epochs for abdominal CT scan samples. Pancreas, kidney, and spleen ROIs are gradually emphasized compared to other parts of the images. Figure taken from [49].

features might be relevant for final segmentation. Using AGs, filters are applied differently depending on relevance, making the network focus on targeted areas while suppressing feature activation in unrelated areas. This operation results in skipping redundant features, and thus, using fewer filters. Figure 2.4 shows the effect of attention coefficients in different epochs of the training process. It can be seen that in initial epochs, most parts of the two images are colored, which indicates the initial uniform distribution of AGs. As the number of epochs increases, from leftmost to rightmost images, only specific parts of images are colored. These areas, red color, in particular, are within ROI boundaries, meaning that features are filtered to pass at specific spatial locations that indicate an organ ROI.

The *Dense-VNet* model, inspired by *V-Networks* [52, 53], used another additive attention mechanism to utilize the shape information for multiple organ segmentation in abdominal CT scans [5]. In order to capture the shape information, they used auxiliary 72^3 learnable parameters to explicitly capture the spatial prior, a low-resolution *3D* map that was upsampled to the segmentation resolution and added as an output of *V-Network*.

Authors of [54] utilized more than one type of attention module. They used a spatial attention module [55] to concentrate on the foreground, a channel-wise attention module to weigh the importance of features at the channel level, and a scale attention module to identify the most relevant features at a multi-scale

level.

Following *U-Net* as a backbone architecture, the work in [6] introduced the residual attention mechanism [56] into *3D* volumes of liver CT scans. The network consisted of three subnetworks in which the first network used *2D* volumes with *2D* residual attention to find the coarse liver boundary. Then, they used the other subsequent subnetworks with *3D* volumes and *3D* residual attention mechanisms to find the exact liver and tumor ROIs, respectively. Finally, the work in [19] proposed an auto-context network to utilize self-supervision contour attention for exploiting high-level residual estimation to obtain the shape prior for liver segmentation.

The other group employed contextual, shape, and contour information to design a multi-task network or embedded it in a loss function. They typically defined contour and shape-related auxiliary tasks and learned them with the main task of ROI segmentation. The DCAN architecture was one of the first examples with a single encoder to learn shared low-level semantic representation from an image and two decoder branches to simultaneously learn ROI segmentation and ROI boundary estimation from this representation [7]. Contour information was used in a contextual network in [18] that utilized a pyramid edge detection module, multi-task module, and interactive attention module to propagate the salient context information. Similarly, the work in [57] used a semantic branch for capturing the semantic information using a spatial attention module and a detail branch to focus more on the contour information in the shallow layers.

In order to extract the boundary information, [8] added a shape-aware loss by comparing the predicted distance maps with the ground truth distance maps and adding the shape information of distance maps into the Dice loss. The distance was defined as the distance between each pixel from the boundary, and the distance maps were obtained from the segmentation map. Thus, the network strictly learned the mapping between the segmentation map and the distance map consisting of spatial information. Authors of [58] also added the boundary information in the form of a loss function comprised of region-based and shape-aware terms. The region-based term maximized the output probability by considering

the overlap of the objects, and the shape-aware term utilized the sign distance and targets both sides of boundaries to have the maximum probability difference.

Finally, the work in [59] exploited both a shape-aware attention module and distance transformation for capturing the boundary information on cardiac CT scans to learn the boundary-aware features from the shape-aware features. Both contour and distance transform maps were used and estimated simultaneously in a V-transition [19] module, then fused with the output of the backbone architecture (*U-Net-like*), and fed into the shape-aware attention module before the final output. The V-transition was composed of a single encoder-decoder and channel-wise attention.

Despite the promising results of these networks on different medical image segmentation problems, none of them used fractal geometry to construct their attention mechanisms or define their auxiliary tasks. Neither the FD had been used as a shape prior information for these networks. Additionally, these studies utilized a feature space in the Euclidean geometry. However, complex texture and boundary structures in an ROI may not follow the smoothness constraint in the Euclidean geometry. Therefore, our approach aims for more accurate segmentation by utilizing fractal geometry through a multi-task network design.

Chapter 3

Methodology

The proposed *MTFD-Net* method relies on 1) quantifying the texture complexity and variation through calculating a map of fractal dimensions (FD map) and 2) designing and training a multi-task network that utilizes this FD map to provide auxiliary supervision to the main task of segmentation. The following sections provide design and implementation details.

3.1 Fractal Dimension Calculation Through Box Counting

The box-counting method was introduced in [60]. It has been widely used for FD estimation as it can be properly applied to structures with and without self-similarity while being easy to implement and more robust. Consider a bounded fractal set S in the Euclidean space. Then, the box-counting method starts with embedding equally sized boxes on an evenly spaced grid such that the boxes cover the whole space. Here the aim is to find the number of boxes consisting of at least one point of interest in that space. The side length of boxes is set to r. This process is repeated with smaller size boxes to see how this number changes by going to finer-scale details. Then, the logarithm of the number of

boxes ($\log N(r)$) versus the logarithm of the scaling factor ($\log 1/r$) is plotted, called Richardson's plot, at each stage. Finally, the FD is derived from repeating this process infinitely and taking the limit from the *log-log* diagram, which is given in Equation 3.1.

$$FD(S) := \lim_{r \to \infty} \frac{\log N(r)}{\log 1/r} \tag{3.1}$$

Given an image space, this equation is slightly different because we cannot resize an image infinitely. Thus, the image space is assumed to be an infinite resolution, and a similar formula is applied to find the FD of images. It is calculated as the slope of the *log-log* diagram by fitting a least square line on the data points. Assuming an $M \times M$ image space where the image is scaled down to $s \times s$ with $1 \leq s \leq M/2$ and $r = s/M$. Then, the FD is calculated as given in Equation 3.2.

$$FD = \frac{logN_r}{log1/r} \tag{3.2}$$

Various box-counting techniques have been proposed and widely used for FD estimation in the literature. The original box-counting method is shown to be effective on binary images. On the other hand, it may not be that effective for calculating the FD of gray-scale images. Therefore, a more efficient approach, so-called differential box counting method was introduced for digital images in 1994 by N. Sarkar and Chaudhuri [39]. In this thesis, we used this method for our FD map generation whose details are given below.

3.2 Fractal Dimension Map Generation

The fractal dimension provides an effective means of quantifying rough surfaces in medical images [61], especially when a surface belonging to an organ or a biological structure exhibits similar patterns at different scales. In this thesis,

we create a fractal dimension (FD) map by separately generating the FD of each pixel in a CT image. To do so, we follow the approach of [14], which uses the differential box-counting method that is known as the most efficient approach for calculating FD maps and works on gray-level intensities [39].

Let I be an image, $I(x, y)$ be the gray-level intensity of a pixel with the coordinates of (x, y), and $w_r(x, y)$ be the kernel with a size of $r \times r$ centered at this pixel. The fractal dimension $FD(x, y)$ of this pixel is estimated in its $R \times R$ neighborhood as follows. First, for the scaling factors $2 \leq r \leq R$, the boxes $N_r(x, y)$ are calculated as given in Equation 3.3, considering the minimum and maximum pixel intensities within a specified kernel $w_r(x, y)$. Then, $FD(x, y)$ is estimated as the slope of a least square regression line on the measurement points of $\log(1/r)$ and $\log(N_r(x, y))$.

$$N_r(x, y) = \frac{R^2}{r^2} \left(\left\lfloor \frac{M_r(x, y) - m_r(x, y)}{r} \right\rfloor + 1 \right) \tag{3.3}$$

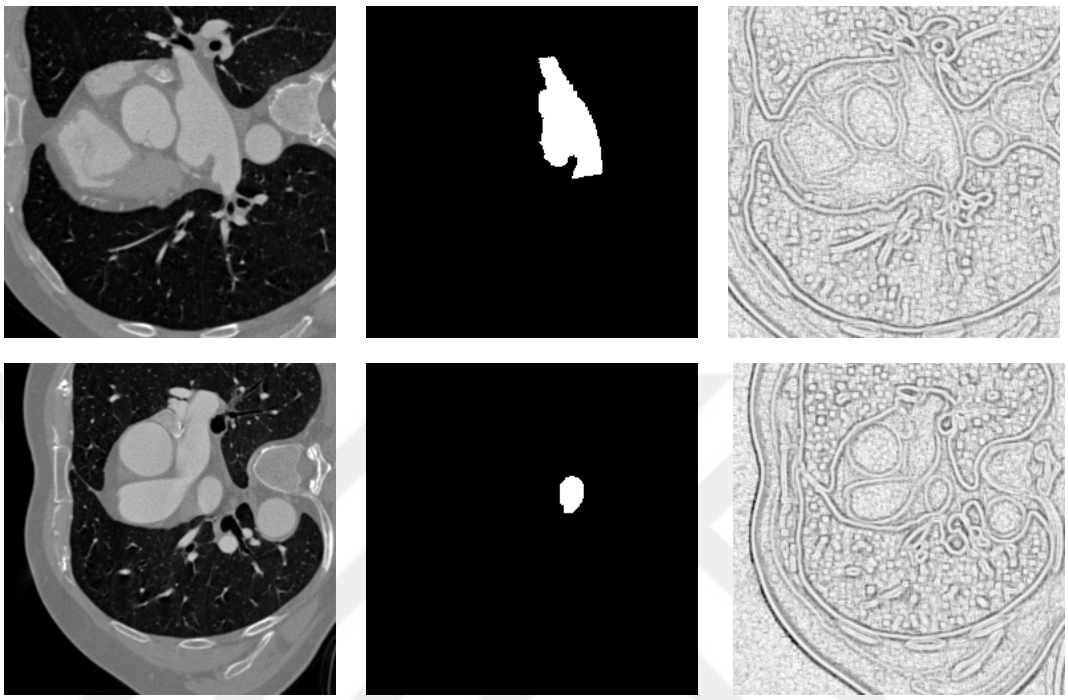
where

$$M_r(x, y) = \max_{(u,v) \in w_r(x,y)} I(u, v), \text{ and} \tag{3.4}$$

$$m_r(x, y) = \min_{(u,v) \in w_r(x,y)} I(u, v). \tag{3.5}$$

In the experiments, R was chosen empirically, considering the resolution of a CT image. When R was selected too small, the kernels did not cover sufficient surrounding pixels to accurately characterize the texture through the fractal dimension. When it was selected too large, the kernels covering very large regions did not bring about additional information. Considering this tradeoff, we selected $R = 7$ in the experiments.

The FD maps generated for two exemplary CT images and their ground truth maps are shown in Figure 3.1. This figure shows that the FD maps provide texture representations complementary to the ground truth maps. Additionally, the fractal dimensions of boundaries between different types of regions exhibit high contrast differences. Learning these FD maps from the original gray-level



(a) (b) (c)

Figure 3.1: (a) Two examples of gray-scale CT images. (b) Ground truths of LA regions. (c) FD maps generated for these images using the differential box-counting method.

CT images provides auxiliary supervision to the network toward the betterment of LA segmentation. The details will be discussed in the next section.

3.3 Multi-Task Network Architecture

A multi-task network includes one main task, and one or more auxiliary tasks that help the network better learn the main task. Such auxiliary tasks are usually related with the main task and provide complementary information to the network while training. A multi-task architecture with N tasks is illustrated in Figure 3.2. There are two types of layers in the network called shared layers (blue blocks) and task-specific layers (pink blocks). The shared layers are a series of convolutional and pooling layers that learn the low-level representation of an input, which corresponds to features common to all tasks. In contrast, task-specific

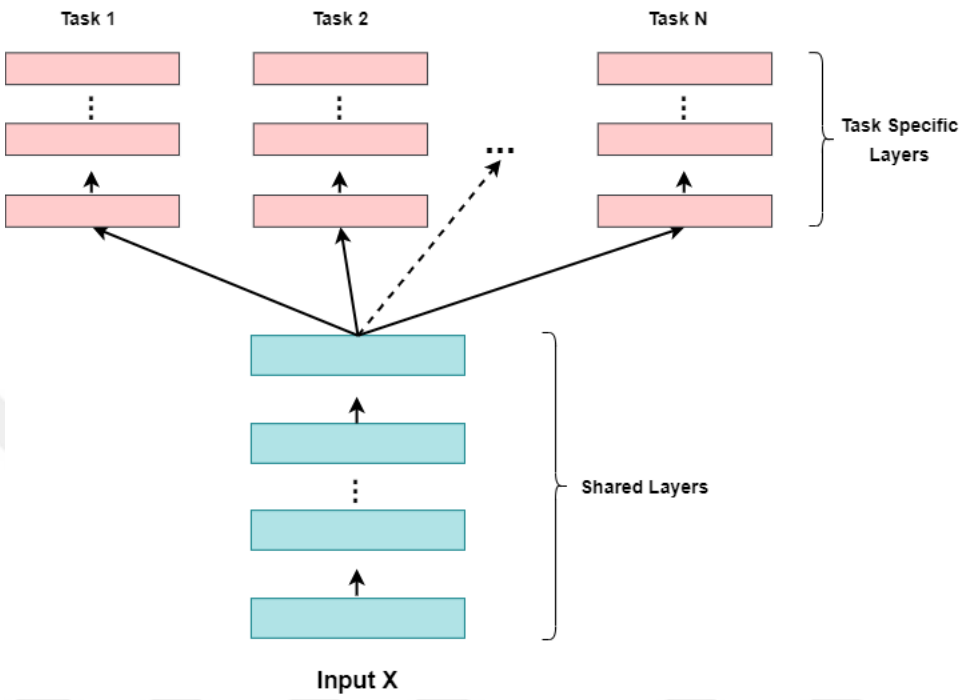


Figure 3.2: A multi-task network. Blue blocks show shared layers that are common hidden representations for the main task and auxiliary tasks. Pink blocks show layers that learn features for each specific task.

layers learn high-level features that embed the learned latent representation from the shared layers to the output layers of the corresponding tasks. The network training objective is a weighted sum of all task-specific losses and the overall loss is minimized through backpropagation.

Multi-task networks have been used for various image segmentation problems. In this thesis, we propose an end-to-end multi-task design that utilizes FD maps for learning the segmentation maps. The proposed *MTFD-Net* model uses a multi-task network to concurrently learn LA segmentation and FD estimation at the same time. The proposed architecture has one shared encoder (down-sampling) path and two separate decoder (up-sampling) paths, one for LA segmentation and the other for FD estimation, which is illustrated in Figure 3.3. The decoder paths decode their output maps from the same feature representations learned at different layers of the shared encoder path. Thus, the encoder needs to keep the necessary context and texture information for the decoders to

accurately realize LA segmentation and FD map estimation, respectively. This concurrent learning is known to be adequate to mitigate the over-fitting problem, which would more likely happen while learning LA segmentation with a single-task network [17].

As given in Figure 3.3, the encoder path includes four blocks of two convolutions and one max pooling. The convolution layers use 3×3 filters followed by the ReLu activation function. The dropout layer with a dropout factor of 0.2 is added at the end of the first convolution as a regularizer to prevent overfitting problem [62]. The max pooling layer uses a 2×2 filter to reduce the spatial dimension by half at the end of each block. This block structure is repeated four times by doubling the number of feature channels at each stage, with 64 channels as the starting point. The fifth block is the bottleneck block with the same two convolution layers without a max pooling layer. This block provides the input to the decoder paths.

The decoder paths include four blocks, each of which consecutively applies the upsampling, concatenation, and two convolution operations. The upsampling operation uses a 2×2 transposed convolution to double the spatial dimension. Its output is concatenated with the features of the corresponding encoder block and then is fed to the convolution operators that are the same as those of the encoder path. The number of feature maps is halved at the end of each decoder block. At the end, 1×1 convolution layers with softmax activation and linear activation functions are applied to obtain the LA segmentation and FD estimation maps, which are the classification and regression tasks, respectively. The details are illustrated in Figure 3.3. As shown in this figure, the overall architecture is a multi-task network that obtains fractal maps through a regression branch and learns fine-detailed segmentation maps through flowing gradient in an end-to-end structure and minimizing the overall loss through backpropagation. Figure 3.4 shows a segmentation map resulting from the network output along with its ground truth and estimated fractal maps.

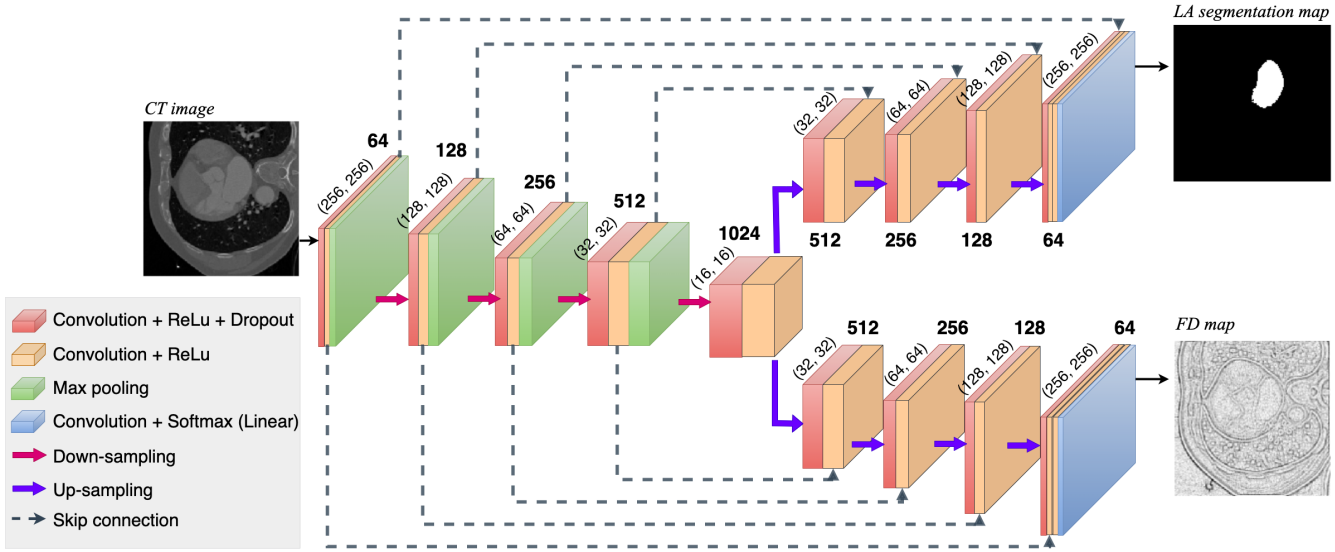


Figure 3.3: Multi-task architecture of the proposed *MTFD-Net*. Each box and arrow corresponds to an operation distinguishable by its color. The numbers (h, w) on the left of each block denote the block's input height and width, respectively. The bold number given for each block is the number of feature maps used by the block's layers.

(a) (b) (c)

Figure 3.4: (a) Ground truth map of a sample test image. (b) Segmentation map predicted by *MTFD-Net*. (c) Its corresponding FD map estimated by the network.

3.4 Implementation Details

All codes were implemented using Python, Matlab, and C. Our network was implemented in a Python3 environment under the open-source deep learning libraries, Tensorflow and Keras. The network was trained on NVIDIA GTX1080Ti GPUs. Evaluation metrics were implemented using Python, Matlab, and C.

The *MTFD-Net* started with an initial filter size of 64 in the encoder path. *ReLu* was used as an activation function in convolutional layers, and thus, we randomly initialized the network using *He* initialization, which was more suitable for *ReLu* [63]. The model was trained to minimize a joint loss function, which was a linear combination of the weighted categorical cross entropy and the mean squared error loss defined for LA segmentation and FD map estimation, respectively. These losses examined each pixel individually and compared it in the predicted segmentation map to the corresponding pixel in the ground truth map. To calculate these losses, the difference between the predictions of our model and the ground truth were calculated and averaged across the entire dataset. The ground truth segmentation maps and the ground truth fractal maps were used to compare the output of our model with these maps in the segmentation task and the regression task, respectively. The contributions of these two losses to the

joint loss function were the same. The pixel contributions in the weighted categorical cross entropy were selected inversely proportional to the class frequencies for balancing ROI compared to the background. A dropout rate of 0.2 was used in the dropout layers. The Adam optimizer [64], with parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$ and the learning rate $2e-5$, was used as an optimizer. The batch size was selected as 1. The training could take a maximum of 600 epochs with an early stopping option. The network training stopped earlier than training on all epochs if there was no improvement in the validation set loss in the last 100 epochs. The network was trained using different configurations and the final optimized configuration was reported.

Learning rates and other network configurations were optimized empirically through experimental trials and fixed for final runs. The network's input was a CT slice where a single input had resolution 256×256. The input images and FD maps were normalized into zero mean and standardized to a variance of one. This was done by subtracting the mean of each slice in a batch and dividing it by its standard deviation. Before training, image slices were randomly shuffled and passed to the network with batch size 1. The network segmentation output was a binary mask, including foreground and background labels. Note that the same configuration was used for the comparison algorithms except that the Adadelta optimizer [65] with learning rate $1e-1$ was used for training a comparison algorithm called *Fourier-Net*.

3.5 Post-Processing

After obtaining the segmentation map of a test set image, we applied post-processing steps on this map. Note that the same post-processing steps were used for the outputs of the comparison algorithms. To do so, we applied erosion followed by small area elimination and dilation to eliminate unnecessary false positive pixels. The details are as follows:

First, an erosion with a structuring element s was applied on the binary segmentation map. A disk kernel with size $(e \times e)$ was used for this purpose where $e \in [3, 5, 7, 9, 11]$ in our grid search. Once erosion was done, we found all connected components in the new map. Second, an area threshold with threshold size h was applied on the eroded map where $h \in [50, 100, 250, 500]$ in the grid search. Finally, a dilation operation with a structuring element $s^{'}$ was performed. Similar to the erosion step, a disk kernel with size $(d \times d)$ was used. In the grid search, $d \in [3, 5, 7, 9, 11]$. The primary purpose of this step was to recover the pixel loss in the first step due to erosion. Additionally, this step and the erosion step might alleviate over-segmentation and under-segmentation errors to some extent, respectively.

Chapter 4

Experiments and Results

We tested the proposed model on an LA dataset provided by a retrospective study involving a cohort of subjects with atrial fibrillation. In order to assess our method, we used various pixel-level metrics, including precision, recall, Dice index, intersection over union (IoU), average symmetric surface distance, and maximum symmetric surface distance.

4.1 Dataset

To evaluate the proposed model and the comparison algorithms, we used the LA segmentation dataset provided by a retrospective study consisting of a cohort of patients with atrial fibrillation. The cohort was obtained by University Hospitals from Iran in 2017-2019. The age of the subjects was in the range of 56-79 years, with an average of 65. The dataset consisted of 2560 gray-scale CT images of 20 subjects (128 images from the CT scan of each subject). The image resolution is 256×256 pixels. The ground truth segmentations were obtained following an approach explained in [66]: initial segmentation maps were created using a 3D region growing algorithm [66] and manual enhancements on these initial maps were performed by a radiologist with six years of experience. The ground truth

maps consisted of two class labels 0 for the background and 1 for the LA region.

The 20 subjects were randomly divided into two halves: CT images of ten subjects were used as the training set and those of the others were used as the test set. The training set was further split into training images (80 percent), on which the network weights were learned by backpropagation, and validation images (20 percent), which were used for early stopping. The images of test subjects were not used in any step of training.

4.2 Evaluation

The LA segmentations were evaluated visually and quantitatively on the test set images. We used six well-known pixel-level metrics for the quantitative assessment of our segmentation results: precision, recall, Dice index, intersection over union, average symmetric surface distance, and maximum symmetric surface distance. Comparing the estimated segmentation and ground truth maps, the number of true positive (TP), false positive (FP), and false negative (FN) pixels were found. Then, the pixel-level metrics were calculated on these numbers.

We will use the ground truth and the segmentation maps shown in Figure 4.1 to illustrate our definitions. In this figure, there are four different regions in the segmented map (S) and the ground truth map (GT). The pixel-level metrics are calculated using:

- ***True Positive (TP)*** represents the number of pixels that are segmented correctly in the segmentation map S and overlap with the pixels in the ground truth map GT. These pixels are shown with green in Figure 4.1.
- ***False Positive (FP)*** represents the number of pixels that were segmented in the segmentation map S but were not a part of pixels in the ground truth map GT. These FP pixels are shown with yellow in Figure 4.1.

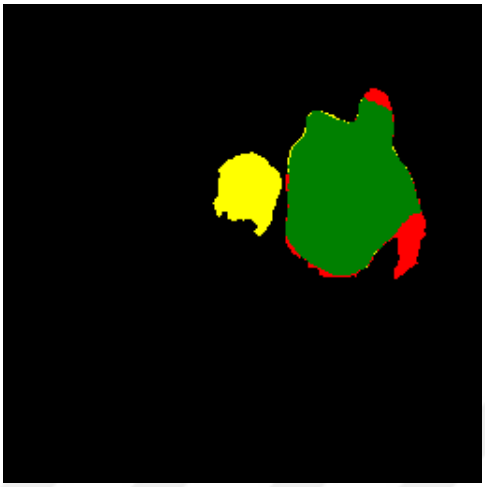


Figure 4.1: An example of the segmentation map and its ground truth. The green area shows the segmented map's overlap with the ground truth, representing true positive (TP) pixels. There are three other main regions representing true negative (TN) pixels with black, false positive (FP) pixels with yellow, and false negative (FN) pixels with red.

- ***True Negative (TN)*** represents the number of pixels that were not segmented in the segmentation map S and did not overlap with the pixels in the ground truth map GT. They are shown with black in Figure 4.1.
- ***False Negative (FN)*** represents the number of pixels that were not segmented in the segmentation map S but were a part of the pixels in the ground truth map GT. These are the red region in Figure 4.1.

Based on these numbers, the precision and the recall are defined as follows:

$$precision = \frac{TP}{TP + FP} \tag{4.1}$$

$$recall = \frac{TP}{TP + FN} \tag{4.2}$$

The Dice index (DSC), which is a trade-off between the precision and the recall, also known as F1-score is defined as follows:

$$DSC = \frac{2\,TP}{2\,TP + FN + FP} \tag{4.3}$$

The next metric is the Jaccard index or intersection over union (IoU). The reason behind adding IoU is that while DSC tends to measure a score closer to the average performance, the IoU score measures something closer to the worst-case performance when an average score over a set of test images is considered. The IoU is calculated as follows:

$$IoU = \frac{TP}{TP + FN + FP} \tag{4.4}$$

These scores are independently calculated for each class (background or foreground) for more accurate performance evaluation. These metrics were reported in two different ways. First, they were calculated on each CT scan of test instances separately and averaged over the test set instances. Second, they were calculated on a volume consisting of all test instances together. These two alternatives are reported separately.

To assess how close the model estimated LA borders to those of the ground truth, two distance-based metrics, the average symmetric surface distance (ASSD) and the maximum symmetric surface distance (MSSD), were calculated as explained in [67]. These two metrics were calculated on the boundary voxels S_B and G_B of the estimated segmentation S and ground truth G volumes of a given 3D CT scan. For each estimated voxel $v \in S_B$, the distance to the closest boundary voxel in the ground truth was calculated. Likewise, for each ground truth voxel, $u \in G_B$, the distance to the nearest boundary voxel in the estimated volume was calculated. The ASSD and MSSD were the average and maximum of all these distances. These metrics were calculated on each CT volume and averaged over the test set subjects. Note that the better segmentations yield higher precision, recall, Dice index, and IoU and lower ASSD and MSSD.

$$d(v, G_B) = \min_{u \in G_B} \|v - u\|^2 \tag{4.5}$$

$$d(u, S_B) = \min_{v \in S_B} \|u - v\|^2 \tag{4.6}$$

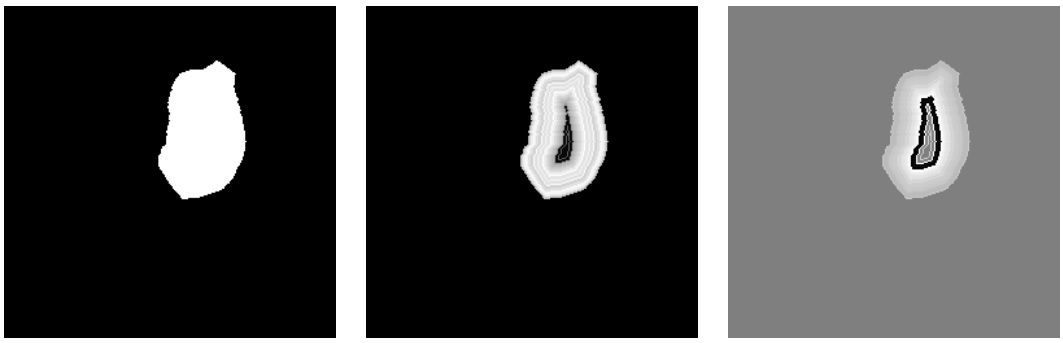
$$ASSD = \frac{\sum_{v \in S_B} d(v, G_B) + \sum_{u \in G_B} d(u, S_B)}{|G_B| + |S_B|} \tag{4.7}$$

$$MSSD = \max\left(\max_{v \in S_B} d(v, G_B), \max_{u \in G_B} d(u, S_B)\right) \tag{4.8}$$

The reason behind using several metrics is that different metrics have their own benefits and drawbacks to show a specific type of error that a model made in the segmentation process.

4.3 Results

The proposed *MTFD-Net* model was compared against two algorithms. The first was the baseline algorithm that used a standard *U-Net* architecture [47]. It was a single-task network with one encoder and one decoder that took a standard normalized CT image as its input and predicted an LA segmentation map as an output. All encoder and decoder operators were the same as those specified in Figure 3.3. This model did not utilize FD maps in any stage of training or as an input. We used this comparison algorithm to understand the effectiveness of using an additional task in network training. The second comparison algorithm (*Fourier-Net*) used a multi-task cascaded FCN network, with precisely the same encoder and decoder architectures shown in Figure 3.3. On the contrary, instead of employing a fractal dimension map, it defined another auxiliary task to utilize shape information in the network design. For that, it extracted Fourier descriptors on the ground truth maps, as explained in [9], and defined the learning of these Fourier descriptors as its auxiliary task. *Fourier-Net* took input images and concurrently learned a segmentation map and Fourier descriptors map by defining a regression and a classification task in its decoder paths, respectively. As opposed to *MTFD-Net*, which used the fractal geometry to model the complexity



(a) (b) (c)

Figure 4.2: (a) Ground truth image. (b) Harmonic amplitude. (c) Harmonic phase.

and variation in the texture of an ROI, the *Fourier-Net* algorithm modeled the shape of the ROI using the Fourier descriptors defined in the Euclidean geometry. Thus, we used this algorithm to compare the effects of using different forms of additional information modeled by the fractal and Euclidean geometries—these two comparison algorithms trained their networks using the same setup with *MTFD-Net*.

The Fourier descriptors have two parts harmonic amplitude and harmonic phase (Figure 4.2). As suggested by [9], we used only a harmonic phase map. The calculated descriptors are N-dimensional. The Fourier descriptors carry more general information in the first dimensions and finer details in the next ones. We experimented with different dimensions in the range of 1 to 10 and chose the highest performance based on validation performances.

4.3.1 Before Post-Processing

The quantitative test set results obtained by the proposed *MTFD-Net* model and the comparison algorithms are presented in Table 4.1. The networks were trained five times for all models, and the average metrics of these five runs and their standard deviations were reported. The table revealed that the proposed *MTFD-Net* model led to higher pixel-level scores (Table 4.1 (a) and Table 4.1

Table 4.1: Test set results obtained by the proposed *MTFD-Net* model and the comparison algorithms. These are the average test set results of the five runs and their standard deviations before applying post-processing. (a) Averaged pixel-level metrics, (b) accumulated pixel-level metrics, and (c) distance-based metrics.

	Averaged			
	Precision	Recall	Dice index	IoU
MTFD-Net	79.65 ± 3.12	91.02 ± 0.90	75.53 ± 2.28	71.36 ± 2.29
Single-task U-Net [47]	64.72 ± 4.41	91.57 ± 1.07	65.38 ± 3.65	60.62 ± 3.66
Multi-task Fourier-Net [9]	67.87 ± 4.26	91.81 ± 1.13	68.02 ± 3.47	63.66 ± 3.49

(a)

	Accumulated			
	Precision	Recall	Dice index	IoU
MTFD-Net	89.21 ± 1.33	89.15 ± 1.19	89.17 ± 0.17	80.45 ± 0.27
Single-task U-Net [47]	82.81 ± 1.77	89.48 ± 1.42	86.00 ± 0.50	75.43 ± 0.77
Multi-task Fourier-Net [9]	83.33 ± 1.50	90.99 ± 1.41	86.98 ± 0.67	76.96 ± 1.03

(b)

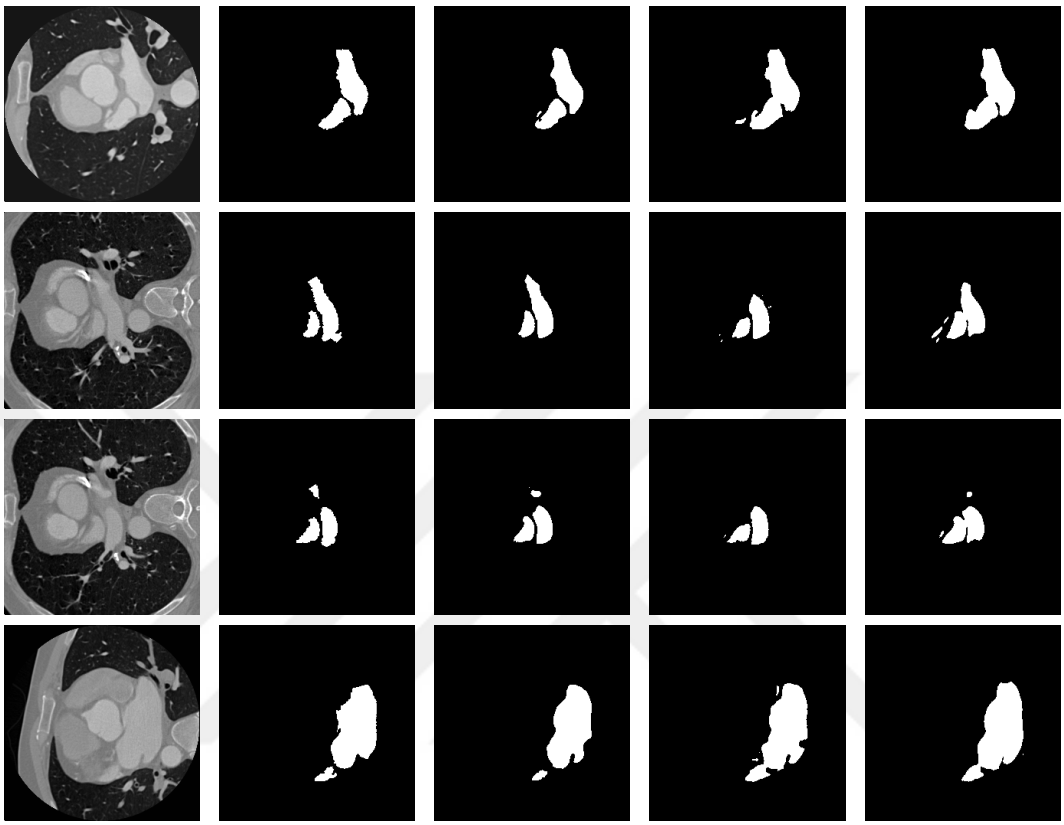
	ASSD	MSSD
MTFD-Net	3.07 ± 0.28	71.52 ± 6.52
Single-task U-Net [47]	4.63 ± 0.37	94.98 ± 6.80
Multi-task Fourier-Net [9]	4.47 ± 0.57	89.28 ± 19.74

(c)

(b)) and lower distance-based metrics (Table 4.1 (c)) than both of the comparison algorithms. As can be seen, our network has outperformed the other two models concerning all six metrics, which also proves the robustness of our approach to a variety of metrics. The first set of pixel-level scores was calculated in two different ways. The averaged scores are those calculated on each test image and averaged over all images. The accumulated scores are those calculated considering all test images together as a volume. To determine whether our model performed significantly better than the comparison models, we applied paired-sample t-test on the Dice scores of our model and the other two models. The t-test ($p < 0.05$) indicated that there was a statistically significant difference between *MTFD-Net* and the other two models. Therefore, *MTFD-Net* resulted in statistically better performance than the comparison algorithms. The distance-based metrics were calculated over each test subject and averaged over all subjects.

These quantitative results were also consistent with visual results shown in Figures 4.3, 4.4, and 4.5. As seen in Figure 4.3, *MTFD-Net* more successfully delineated the boundaries of adjacent LA regions compared to other algorithms. Additionally, when there is only one LA region in the image, *MTFD-Net* led to better-shaped estimations (Figure 4.4), even though the *Fourier-Net* comparison algorithm explicitly modeled the shape information of an ROI. Finally, as seen in Figure 4.5, all algorithms might fail for some images, especially when the size of an ROI is very small compared to the background (second row). Even on these images, we observed that *MTFD-Net* outperformed the other algorithms. In the first row of Figure 4.5, *U-Net* resulted in some over-segmentation while leaving out the necessary part, and *Fourier-Net* resulted in under-segmentation. In contrast, our model still resulted in a segmentation map closest to the ground truth. We had the same observation for the image shown in the second row of Figure 4.5.

In the experiments, it was necessary to find the most suitable number of Fourier coefficients N for the *Fourier-Net* algorithm that gave the best performed. To this end, three runs were performed on five values of N, and the pixel-level metrics were calculated for the test set and validation set images. Then, the number N that led to the highest average validation score was selected, and the final five



(a) (b) (c) (d) (e)

Figure 4.3: (a) Example test set images. (b) Ground truths. Results of (c) the proposed *MTFD-Net*, (d) the single-task *U-Net* [47], and (e) the multi-task *Fourier-Net* [9] algorithms. These visual results are obtained before post-processing.

runs for comparison with *MTFD-Net* were performed on the selected N. Table 4.2 shows the scores for different values of $N = \{1, 2, 3, 5, 10\}$. It can be seen that the highest validation Dice score was obtained when N was set to 3. Thus, $N = 3$ was used in the experiments.

4.3.2 After Post-Processing

This section reports the quantitative and qualitative results after applying post-processing. After post-processing, our model also outperformed the other two models. Table 4.3 and Figures 4.6, 4.7, and 4.8 show the quantitative and visual

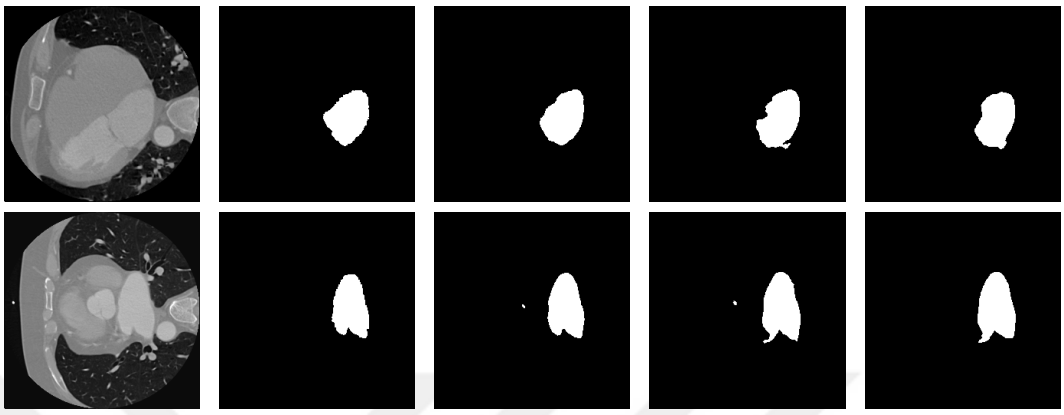
Table 4.2: *Fourier-Net* scores for different N values on (a) test set and (b) validation set. Averages and standard deviations are shown for three runs before applying post-processing.

	Test Set			
	Precision	Recall	Dice index	IoU
N=1	81.36 ± 3.59	91.90 ± 2.36	86.18 ± 1.20	75.74 ± 1.84
N=2	82.80 ± 2.09	91.48 ± 1.89	86.89 ± 0.86	76.84 ± 1.36
N=3	83.36 ± 2.07	90.48 ± 1.07	86.76 ± 0.77	76.62 ± 1.21
N=5	85.47 ± 0.94	90.33 ± 0.31	87.83 ± 0.44	78.30 ± 0.69
N=10	86.71 ± 1.45	89.03 ± 2.15	87.89 ± 0.77	78.40 ± 1.22

(a)

	Validation Set			
	Precision	Recall	Dice index	IoU
N=1	90.74 ± 2.09	93.50 ± 1.83	92.05 ± 0.14	85.27 ± 0.24
N=2	91.62 ± 1.33	93.47 ± 2.37	92.51 ± 0.50	86.07 ± 0.86
N=3	91.85 ± 0.90	93.34 ± 1.59	92.52 ± 0.44	86.08 ± 0.76
N=5	91.88 ± 0.69	93.11 ± 0.69	92.32 ± 0.26	86.03 ± 0.64
N=10	92.98 ± 1.00	91.83 ± 2.03	92.39 ± 0.55	85.84 ± 0.93

(b)



(a) (b) (c) (d) (e)

Figure 4.4: (a) Example test set images. (b) Ground truths. Results of (c) the proposed *MTFD-Net*, (d) the single-task *U-Net* [47], and (e) the multi-task *Fourier-Net* [9] algorithms. These visual results are obtained before post-processing.

results, respectively. In the images shown in Figure 4.6, it can be seen that post-processing improved the results of each network visually. This post-processing was effective in eliminating false positives that were detected as an ROI around the real ROI.

In the next set of images shown in Figure 4.7, it can be seen that post-processing might decrease segmentation accuracy for all the networks. This mostly happened when there were small or multiple ROIs within a segmentation map. In the first row of Figure 4.7, the ROIs lost their original shape, which was closer to the ground truth, while in the second row, one or the whole part of the ROI, and in the third row, the real ROI was eliminated from the segmentation map. In the last row, there was a small tail in the bottom right corner of the ROI in each segmentation map, which was eliminated after post-processing. This might be negligible for the overall quantitative scores, but it visually changed the original appearance, which was close to the ground truth. The last set of images shown in Figure 4.8 represented the cases where post-processing affected models differently. For instance, it can be seen that the segmentation maps of *Fourier-Net* after post-processing were less negatively affected by post-processing compared to the other two models.

Table 4.3: Test set results obtained by the proposed *MTFD-Net* model and the comparison algorithms. These are the average test set results of the same five runs and their standard deviations after applying post-processing. (a) Averaged pixel-level metrics, (b) accumulated pixel-level metrics, and (c) distance-based metrics.

	Averaged			
	Precision	Recall	Dice index	IoU
MTFD-Net	95.48 ± 0.96	85.49 ± 0.84	85.27 ± 0.32	82.22 ± 0.31
Single-task U-Net [47]	91.14 ± 1.99	86.56 ± 0.90	83.76 ± 1.00	80.08 ± 0.97
Multi-task Fourier-Net [9]	94.78 ± 0.93	84.70 ± 0.84	83.43 ± 0.38	80.46 ± 0.36

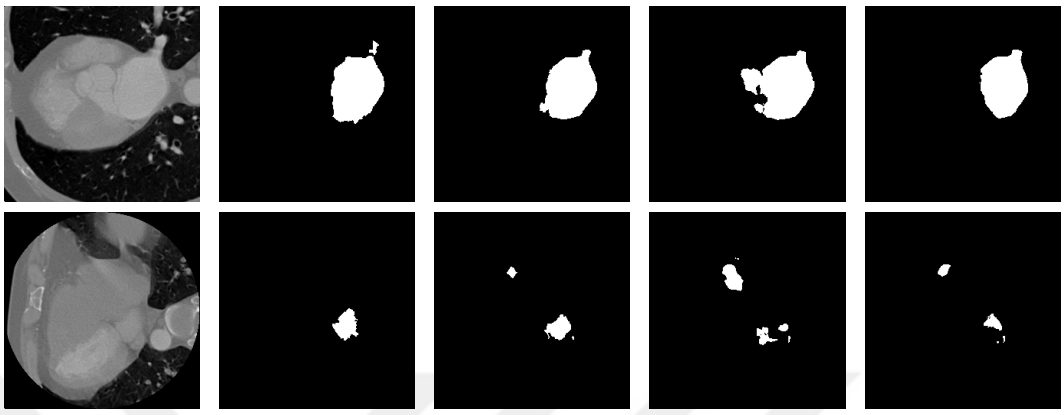
(a)

	Accumulated			
	Precision	Recall	Dice index	IoU
MTFD-Net	92.79 ± 1.18	87.04 ± 1.32	89.81 ± 0.23	81.50 ± 0.42
Single-task U-Net [47]	88.88 ± 1.15	87.25 ± 1.50	88.04 ± 0.51	78.64 ± 0.80
Multi-task Fourier-Net [9]	89.39 ± 1.09	87.72 ± 1.38	88.73 ± 0.72	79.75 ± 1.17

(b)

	ASSD	MSSD
MTFD-Net	2.23 ± 0.10	30.15 ± 3.92
Single-task U-Net [47]	2.98 ± 0.51	38.14 ± 4.58
Multi-task Fourier-Net [9]	14.47 ± 0.96	73.91 ± 2.87

(c)



(a) (b) (c) (d) (e)

Figure 4.5: (a) Example test set images. (b) Ground truths. Results of (c) the proposed *MTFD-Net*, (d) the single-task U-Net [47], and (e) the multi-task *Fourier-Net* [9] algorithms. These visual results are obtained before post-processing.

Additionally, Figures 4.9 and 4.10 depicted the 3D LA construction for two exemplary test set subjects. These were constructed by the ITK-snap tool [68] on the ground truths and the estimated segmentation maps. These figures showed that the LA region estimations of the *MTFD-Net* model led to more accurate constructions than the comparison algorithms.

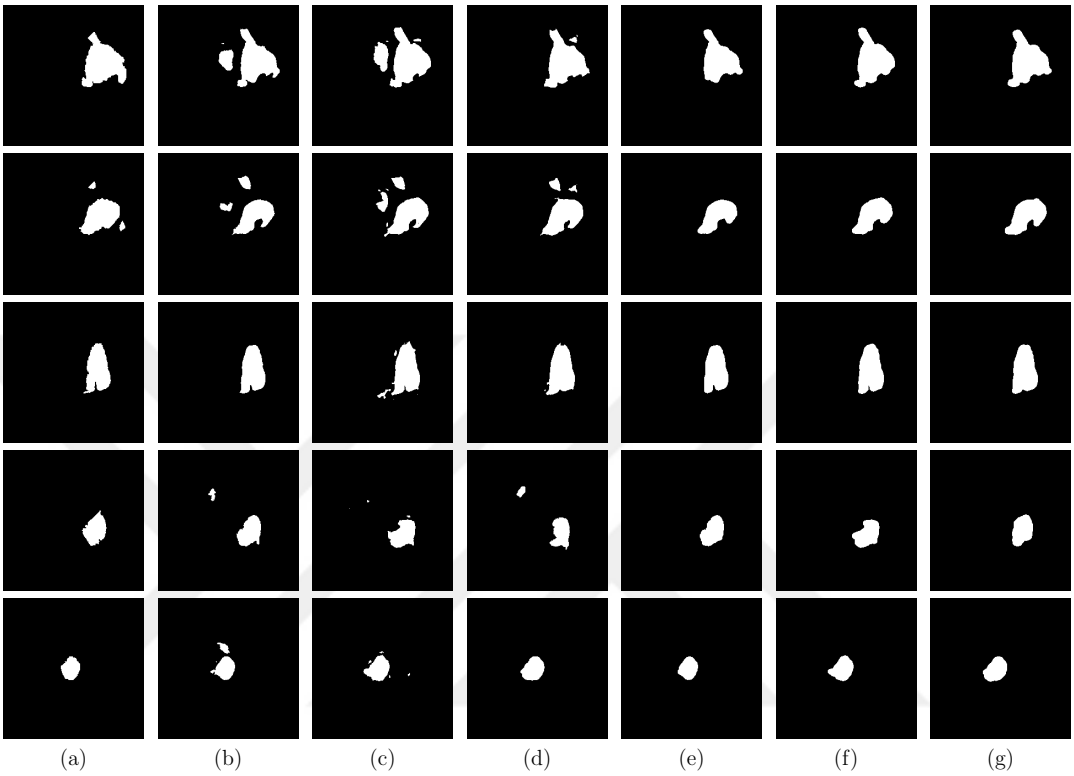


Figure 4.6: (a) Example ground truths. Results of (b) the proposed *MTFD-Net*, (c) the single-task *U-Net* [47], and (d) the multi-task *Fourier-Net* [9] algorithms before post-processing. Results of (e) the proposed *MTFD-Net*, (f) the single-task U-Net [47], and (g) the multi-task *Fourier-Net* [9] algorithms after post-processing.

In summary, comparing our proposed model with the single-task *U-Net* algorithm, the experimental results indicated the effectiveness of learning shared feature representations from multiple tasks for LA segmentation, which is indeed known to be effective for many domains [17]. Comparing it with *Fourier-Net* [9], the experiments demonstrated that for LA segmentation, modeling the texture variations in an ROI using the fractal geometry was more effective than modeling the shape information of the ROI using the Euclidean geometry.

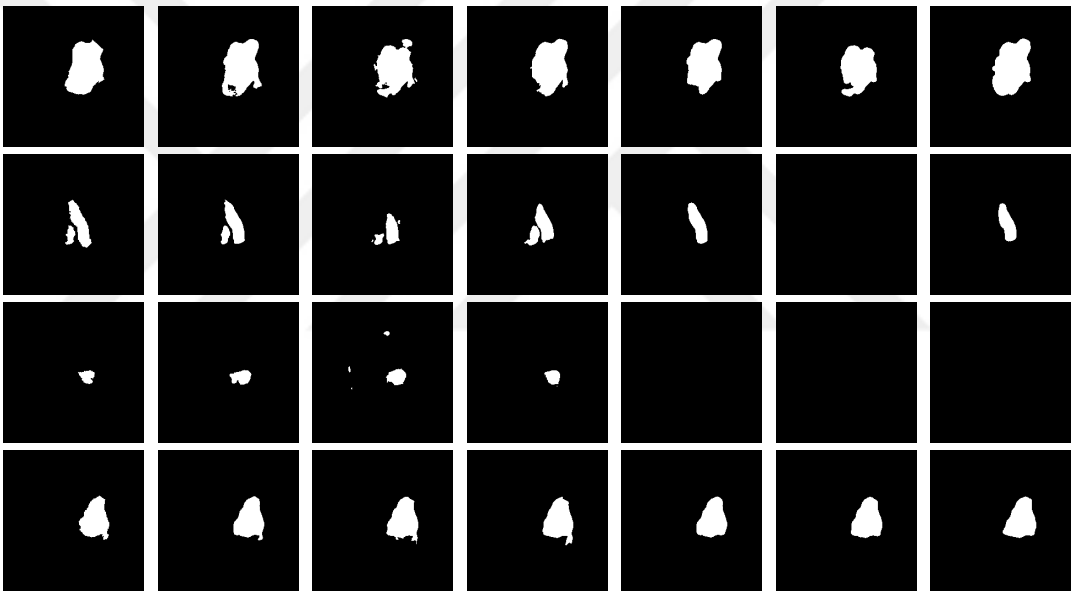


Figure 4.7: (a) Example ground truths. Results of (b) the proposed *MTFD-Net*, (c) the single-task *U-Net* [47], and (d) the multi-task *Fourier-Net* [9] algorithms before post-processing. Results of (e) the proposed *MTFD-Net*, (f) the single-task U-Net [47], and (g) the multi-task *Fourier-Net* [9] algorithms after post-processing.

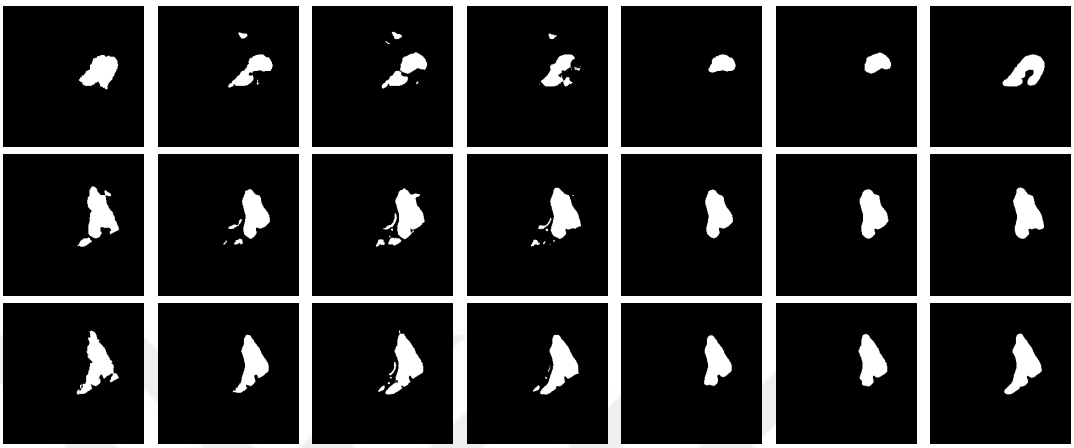


Figure 4.8: (a) Example ground truths. Results of (b) the proposed *MTFD-Net*, (c) the single-task *U-Net* [47], and (d) the multi-task *Fourier-Net* [9] algorithms before post-processing. Results of (e) the proposed *MTFD-Net*, (f) the single-task U-Net [47], and (g) the multi-task *Fourier-Net* [9] algorithms after post-processing.

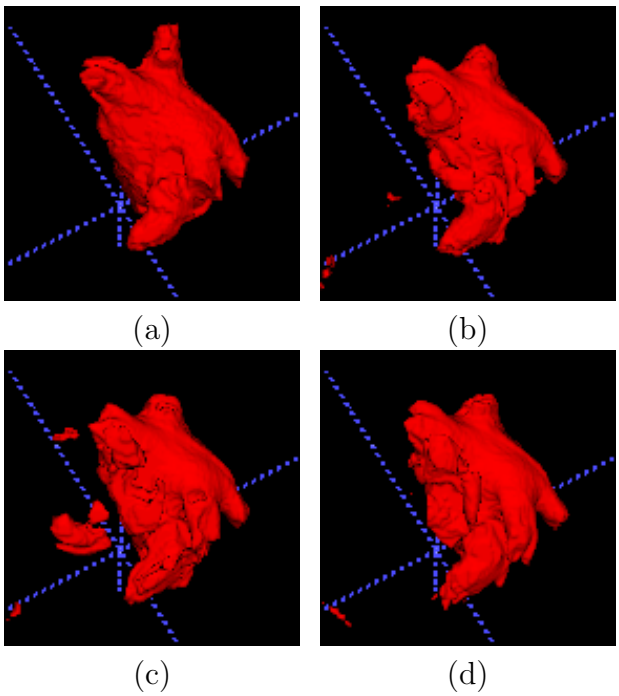


Figure 4.9: For an exemplary test set subject, LA construction (a) from the 3D volume of the ground truth maps, and from the 3D volume of the segmentation maps generated by (b) the proposed *MTFD-Net*, (c) the single-task *U-Net* [47], and (d) the multi-task *Fourier-Net* [9] algorithms.

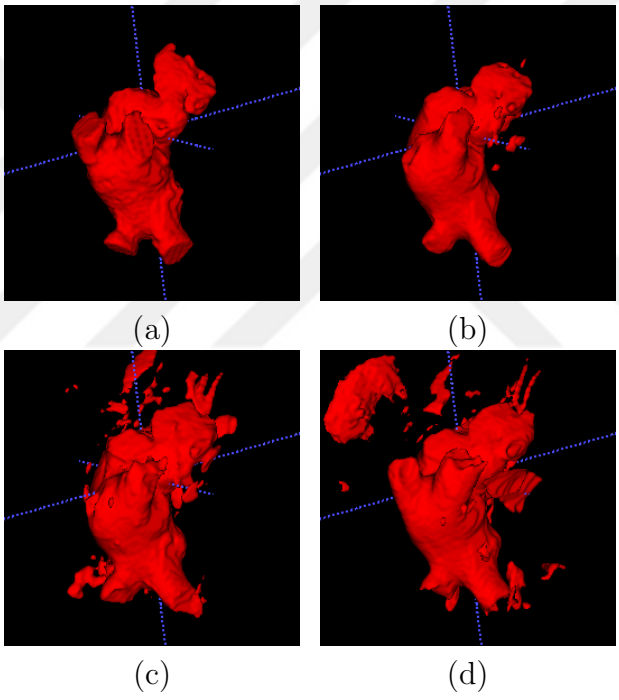


Figure 4.10: For another exemplary test set subject, LA construction (a) from the 3D volume of the ground truth maps, and from the 3D volume of the segmentation maps generated by (b) the proposed *MTFD-Net*, (c) the single-task *U-Net* [47], and (d) the multi-task *Fourier-Net* [9] algorithms.

Chapter 5

Conclusion

This thesis proposed a novel approach that utilized the concept of fractal dimension for LA segmentation and benefited from learning the complexity of LA texture through a deep end-to-end convolutional network. We experimented with the LA task on a CT heart dataset. The visual and quantitative results were reported in the previous chapter. First, FD maps for each slice were generated from the original CT images. To this end, a variation of the differential box-counting method was used to extract the pixel-wise FD on the entire CT slice, producing a map with the same resolution as the original slice. These maps were utilized in an auxiliary task in a multi-task network design to help better learn binary segmentation maps simultaneously. Through our experiments on the LA dataset, it was illustrated that our proposed *MTFD-Net* provided substantially superior performance compared to the well-known networks single-task *U-Net* and multi-task shape-preserving method, *Fourier-Net*.

The superiority of our model was shown both quantitatively and qualitatively. Based on visual results, three significant advantages of our approach were concluded. First, it separated the boundary sides better than the other two networks. Second, it yielded less under-segmentation and over-segmentation. Third, even though *Fourier-Net* was a shape-preserving architecture that utilized the prior

shape information, we demonstrated cases where our model resulted in a better shape LA ROI, closest to the ground truth. Moreover, most of the existing research only uses MRI scans for their segmentation approaches, which provide isolated angle views for cardiac structures, making the segmentation easier. On the other hand, we used CT scans, which have higher intensity variance among different slices.

As future work, there might be two possible extensions to our method. We conducted preliminary experiments to explore these possibilities. First, *MTFD-Net* was tested on an LA task. To determine the effectiveness of our method for segmenting other complex structures surrounding the LA, we conducted experiments on the pulmonary vein segmentation task. Pulmonary veins join the LA entering the posterior part of the LA. Like LA, pulmonary vein anatomy is complex due to its variable morphology, ostial diameter, and orientation. Moreover, its varying size makes segmentation harder, especially in ROIs containing several separated small parts.

We applied our approach to the pulmonary vein segmentation task on a similar dataset to the LA dataset. The same University Hospitals also provided this dataset from Iran in 2017-2019. The dataset consisted of the same patients having a new resolution of 512×512 where pulmonary veins were targeted. The dataset consisted of 3743 gray-scale CT images of 20 subjects (varying volume sizes for the subjects). The ground truth maps were generated using the method in [66]. The ground truth maps were labeled as 0 for the background and 1 for the pulmonary vein region. The subjects were divided into two halves of training and test sets. The division criterion was to split all CT slices into two halves since each subject had a varying volume size. To this end, each patient's volume size was taken into account while randomly splitting the subjects. The training set was further divided into 80 percent for the training set and 20 percent for the validation set. The test set subjects were not involved in any stage of training the models.

We trained *MTFD-Net* on this dataset to segment pulmonary vein ROI and obtained preliminary results similar to *Fourier-Net* and significantly better than the single-task network *U-Net*. Table 5.1 presented the average pixel-level scores

Table 5.1: Test set results obtained by the proposed *MTFD-Net* model and the comparison algorithms on pulmonary vein segmentation task. These are the averaged pixel-level metrics for average test set results of the three runs and their standard deviations before applying post-processing.

	Averaged			
	Precision	Recall	Dice index	IoU
MTFD-Net	76.75 ± 1.74	88.79 ± 0.36	82.32 ± 0.75	69.95 ± 1.09
Single-task U-Net [47]	64.42 ± 3.08	89.50 ± 0.84	74.88 ± 1.87	59.87 ± 2.40
Multi-task Fourier-Net [9]	77.87 ± 0.65	86.15 ± 3.27	82.43 ± 0.17	70.12 ± 0.24

for three runs on *MTFD-Net* and the comparison models. This table revealed that our model outperformed the single-task *U-Net* and obtained very similar scores to the multi-task shape preserving model, *Fourier-Net*. Since the pixel-level scores of our model were very close to *Fourier-Net*, we applied the paired-sample t-test on the Dice scores of the two models. The t-test showed that there was no statistically significant difference between *MTFD-Net* and *Fourier-Net* Dice scores ($p = 0.77 > 0.05$). Thus, *Fourier-Net* did not perform better even though it preserves the shape information and appears to be the best method for pulmonary vein segmentation due to the high variability in the shape and size of veins. We believe that purposefully improving *MTFD-Net* network parameters or training on a new network design can enhance the performance of our technique. The visual results were consistent with the quantitative results (see Figures 5.1, 5.2, and 5.3). The investigation of improving the *MTFD-Net* model for pulmonary vein segmentation is considered as the first future work.

One may develop a new complex single-task, multi-task, or cascaded architecture that can use FD maps more effectively for segmentation. For example, one can use the FD maps in the input layer as its calculation does not require the ground truth maps. Likewise, their estimation can be defined as an intermediate task in a cascaded design. This is the second future research direction of this Thesis.

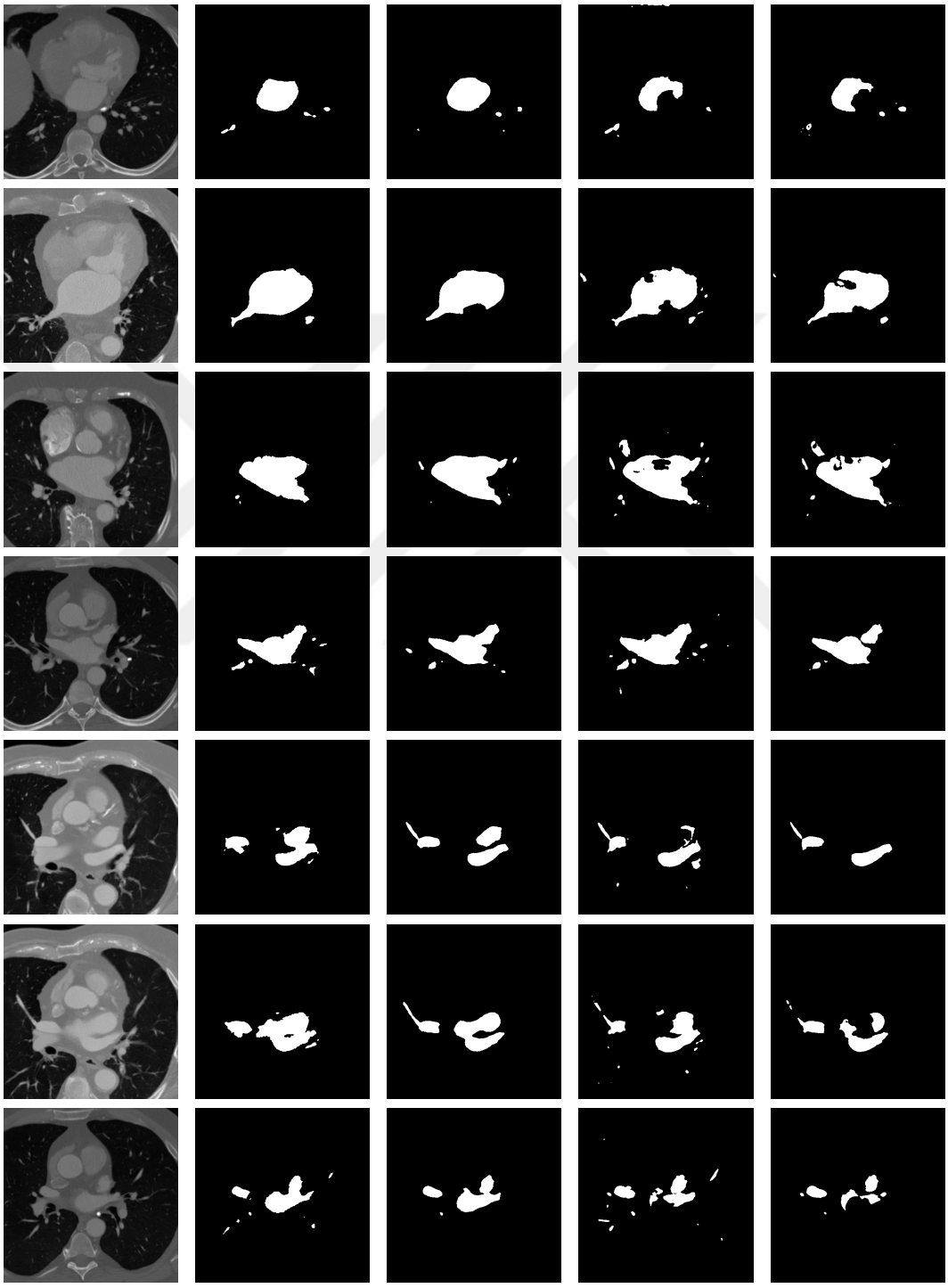
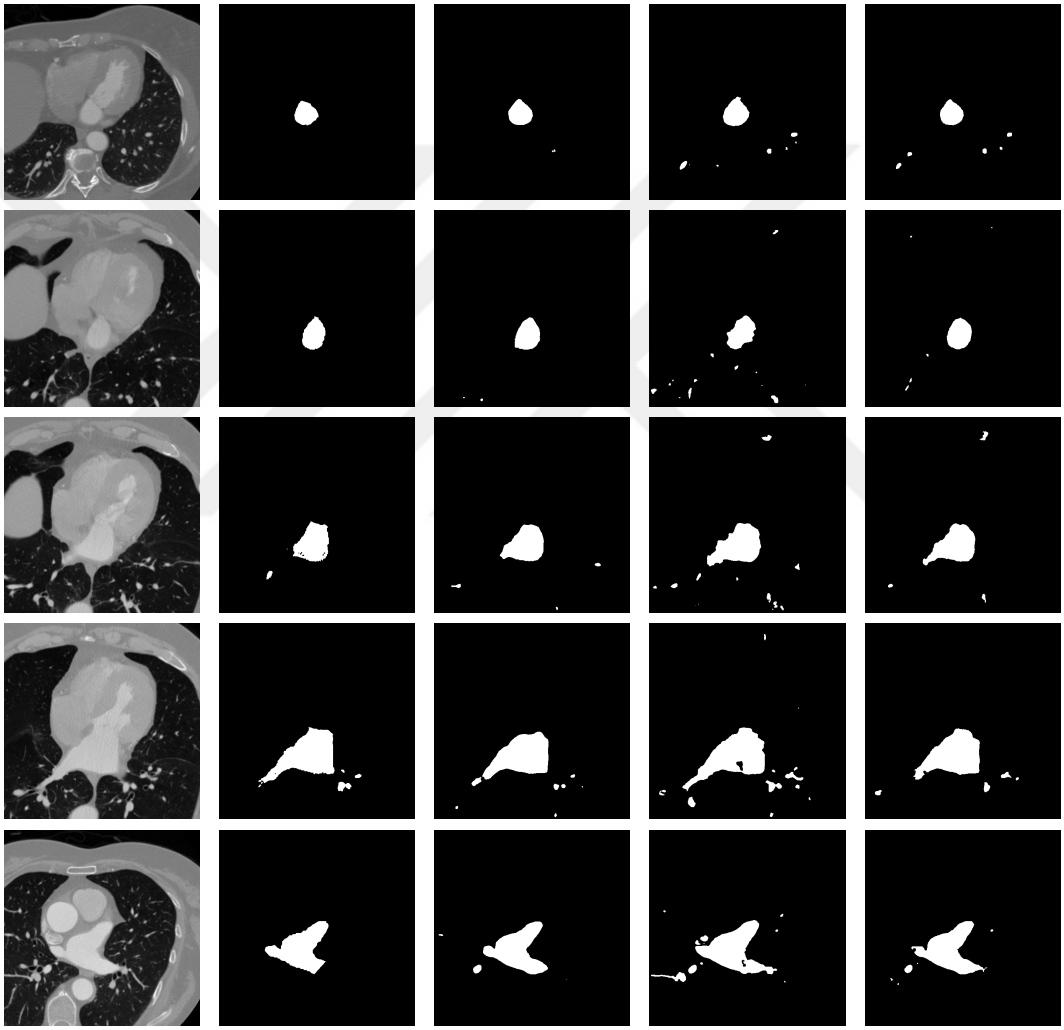
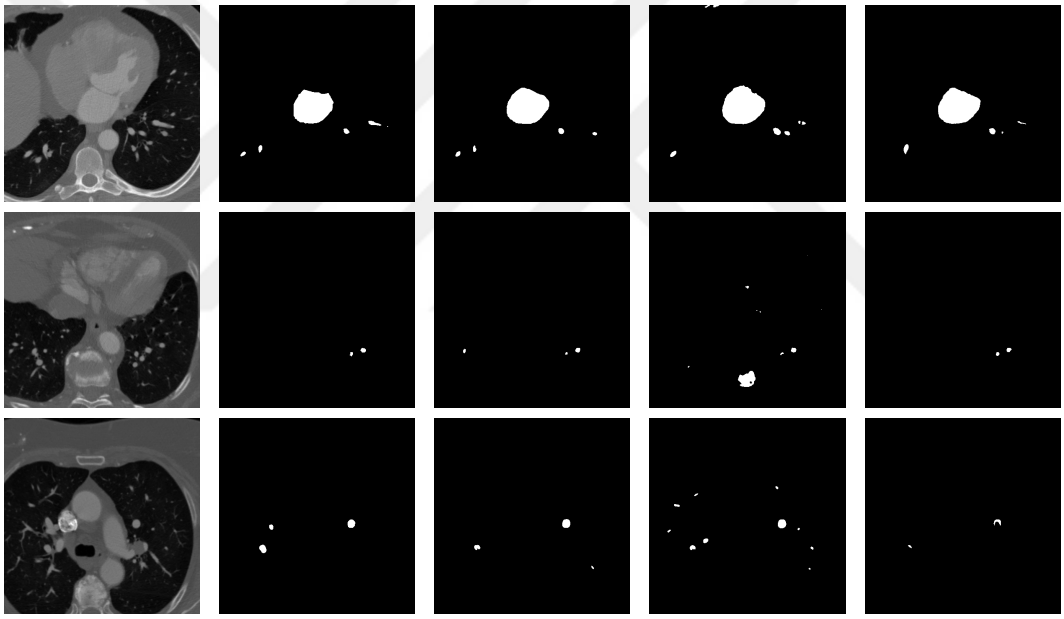


Figure 5.1: Visual results on pulmonary vein segmentation task. (a) Example test set images. (b) Ground truths. Results of (c) the proposed *MTFD-Net*, (d) the single-task *U-Net* [47], and (e) the multi-task *Fourier-Net* [9] algorithms.



(a) (b) (c) (d) (e)

Figure 5.2: Visual results on pulmonary vein segmentation task. (a) Example test set images. (b) Ground truths. Results of (c) the proposed *MTFD-Net*, (d) the single-task *U-Net* [47], and (e) the multi-task *Fourier-Net* [9] algorithms.



(a) (b) (c) (d) (e)

Figure 5.3: Visual results on pulmonary vein segmentation task. (a) Example test set images. (b) Ground truths. Results of (c) the proposed *MTFD-Net*, (d) the single-task *U-Net* [47], and (e) the multi-task *Fourier-Net* [9] algorithms.

Bibliography

[1] Z. Malchano *et al.*, "Integration of cardiac CT/MR imaging with three-dimensional electroanatomical mapping to guide catheter manipulation in the left atrium: implications for catheter ablation of atrial fibrillation," *J. Cardiovasc. Electrophysiol.*, vol. 17, no. 11, pp. 1221–1229, 2006.

[2] F. T. Laurens *et al.*, " Fusion of Electroanatomical Activation Maps and Multislice Computed Tomography to Guide Ablation of a Focal Atrial Tachycardia in a Fontan Patient," *J. Cardiovasc. Electrophysiol.*, vol. 17, no. 4, pp. 431–434, 2006.

[3] Z. Xiong *et al.*, "A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging," *Med. Image Anal.*, vol. 67, p. 101832, 2021.

[4] B. B. Mandelbrot, "Fractal Geometry of Nature," W. H. Freeman and Co., New York, 1982.

[5] E. Gibson *et al.*, "Automatic multi-organ segmentation on abdominal CT with dense v-networks," *IEEE Trans. Med. Imaging*, vol. 37, no. 8, pp. 1822–1834, 2018.

[6] Q. Jin *et al.*, "RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans," *Front. Bioeng. Biotech.*, p.1475, 2020.

[7] H. Chen, X. Qi, L. Yu, and P. Heng, "DCAN: Deep contour-aware networks for accurate gland segmentation," in *Proc. IEEE Conf. Comp. Vis. Pattern Recognit.*, pp. 2487–2496, 2016.

[8] H. Li *et al.*, "Deep distance map regression network with shape-aware loss for imbalanced medical image segmentation," in *Proc. Int. Workshop on Machine Learning in Medical Imaging*, pp. 231–240, 2020.

[9] S. Cansiz *et al.*, "FourierNet: Shape-preserving network for Henle's fiber layer segmentation in optical coherence tomography images," 2022, arXiv:2201.06435. [Online]. Available: http://arxiv.org/abs/2201.06435.

[10] A. H. Mir *et al.*, "Texture analysis of CT images," *IEEE Eng. Med. Biol. Mag.*, vol. 14, no. 6, pp. 781–786, 1995.

[11] S. R. Nayak *et al.*, "Analysing roughness of surface through fractal dimension: A review," *Image. Vis. Comput.*, pp. 21–34, 2019.

[12] G. G. Medioni *et al.*, "A note on using the fractal dimension for segmentation," *Proc. IEEE Workshop on Computer Vision*, pp. 25-30, 1984.

[13] J. Keller *et al.*, "Texture description and segmentation through fractal geometry," in *Comput. Vis. Graph. Image Process.*, vol. 45, no. 2, pp. 150–166, 1989.

[14] O. S. Al-Kadi and D. Watson, "Texture analysis of aggressive and nonaggressive lung tumor CE CT images," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 7, pp. 1822–1830, 2008.

[15] R. Lopes and N. Betrouni, "Fractal and multifractal analysis: A review," *Med. Image Anal.*, vol. 13, no. 4, pp. 634–649, 2009.

[16] B.B. Chaudhuri and N. Sarkar, "Texture segmentation using fractal dimension," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 1, pp. 72–77. 1995.

[17] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, pp. 41–75. 1997.

[18] R. Wang, S. Chen, C. Ji, J. Fan, and Y. Li, "Boundary-aware context neural network for medical image segmentation," *Med. Image Anal.*, vol. 78, p. 102395, 2022.

[19] M. Chung, J. Lee, J. Lee, and Y.-G. Shin, "Liver segmentation in abdominal CT images via auto-context neural network and self-supervised contour attention," *Artif. Intell. Med.*, vol. 113, p. 102023, 2020.

[20] B. B. Mandelbrot *et al.*, "How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension," *Science*, vol. 156, no. 3775, pp. 636–638, 1967.

[21] W. L. Lee *et al.*, "A robust algorithm for the fractal dimension of images and its applications to the classification of natural images and ultrasonic liver images," *Signal Process.*, vol. 90, no. 6, pp. 1894–1904, 2010.

[22] W. S. Chen *et al.*, "Two algorithms to estimate fractal dimension of gray-level images," *Opt. Eng.*, vol. 42, no. 3, pp. 2452–2464, 2003.

[23] P. Asvestas *et al.*, "A power differentiation method of fractal dimension estimation for 2-D signals," *J. Vis. Commun. Image Represent.*, vol. 9, no. 4, pp. 392–400, 1998.

[24] P. Nicolini *et al.*, "Hausdorff dimension of a particle path in a quantum manifold," *Phys. Rev. D*, vol. 83, no. 2, p. 024017, 2011.

[25] I. Pilgrim *et al.*, "Fractal analysis of time-series data sets: Methods and challenges," *Fractal Analysis*, 2018.

[26] S. Kido *et al.*, "Fractal analysis of internal and peripheral textures of small peripheral bronchogenic carcinomas in thin-section computed tomography: Comparison of bronchioloalveolar cell carcinomas with nonbronchioloalveolar cell carcinomas," *J. Comput. Assist. Tomogr.*, vol. 27, no. 1, pp. 56–61, 2003.

[27] W. L. Lee *et al.*, "Ultrasonic liver tissue classification by fractal feature vector based on M-band wavelet transform," *IEEE Trans. Med. Imaging*, vol. 22, no. 3, pp. 382–392, 2003.

[28] C. M. Wu *et al.*, "Multithreshold dimension vector for texture analysis and its application to liver-tissue classification," *Pattern Recognit.*, vol. 26, no. 1, pp. 137–144, 1993.

[29] R. Uppaluri *et al.*, "Fractal analysis of high-resolution CT images as a tool for quantification of lung diseases," *Med. Imaging 1995*, vol. 2433, pp. 133–142, 1995.

[30] D. R. Chen *et al.*, "Classification of breast ultrasound images using fractal feature," *Clin. Imaging*, vol. 29, no. 4, pp. 235–245, 2005.

[31] A. I. Penn *et al.*, "Discrimination of MR images of breast masses with fractal-interpolation function models," *Acad. Radiol.*, vol. 6, no. 3, pp. 156–163, 1999.

[32] M. Biswas *et al.*, "Fractal dimension estimation for texture images: a parallel approach," *Pattern Recognit. Lett.*, vol. 19, no. 3–4, pp. 309–313. 1998.

[33] A. Balghonaim *et al.*, "A maximum likelihood estimate for two-variable fractal surface," *IEEE Trans. Image Process.*, vol. 7, no. 12, pp. 1746–1753, 1998.

[34] J. J. Gagnepain *et al.*, "Fractal approach to two dimensional and three dimensional surface roughness," *Wear*, vol. 109, no. 1–4, pp. 119–126, 1986.

[35] J. M. Keller *et al.*, "Texture description and segmentation through fractal geometry," *Comput. Gr. Image Process.*, vol. 45, no. 2, pp. 150–166, 1989.

[36] X. Jin *et al.*, "A practical method for estimating fractal dimension," *Pattern Recognit. Lett.*, vol. 16, no. 5, pp. 457–464, 1995.

[37] S. Buczkowski *et al.*, "The modified box-counting method: analysis of some characteristics parameters," *Pattern Recognit.*, vol. 31, no. 4, pp. 411–418, 1998.

[38] R.S. Pindyck *et al.*, "Econometric models and economic forecasts," *McGraw-hill*, 1991.

[39] N. Sarkar *et al.*, "An efficient differential box-counting approach to compute fractal dimension of image," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 24, no. 1, pp. 115–120, 1994.

[40] J. Li *et al.*, "An improved box-counting method for image fractal dimension estimation," *Pattern Recognit.*, vol. 42, no. 11, pp. 2460–2469, 2009.

[41] Y. U. Liu *et al.*, "An improved differential box-counting method to estimate fractal dimensions of gray-level images," *J. Vis. Commun. Image Represent.*, vol. 25, no. 5, pp. 1102–1111, 2014.

[42] L. Yu *et al.*, "Coarse iris classification using box-counting to estimate fractal dimensions," *Pattern Recognit.*, vol. 38, no. 11, pp. 1791–1798, 2005.

[43] S. R. Nayak *et al.*, "Analysing Fractal Dimension of Color Images," *International Conference on Computational Intelligence and Networks*, pp. 156–159, 2015.

[44] S. R. Nayak *et al.*, "An improved algorithm to estimate the fractal dimension of gray scale images," *International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, pp. 1109–1114, 2016.

[45] J. Long *et al.*, "Fully convolutional networks for semantic segmentation," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3431–3440, 2015.

[46] K. Simonyan *et al.*, "Very deep convolutional networks for large-scale image recognition," *Pattern Recognit.*, arXiv preprint arXiv:1409.1556, 2014.

[47] O. Ronneberger *et al.*, "U-net: convolutional networks for biomedical image segmentation," *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, vol. 9351, pp. 231–241, 2015.

[48] L. C. Chen *et al.*, "Attention to scale: Scale-aware semantic image segmentation," *Proc. IEEE Conf. Comp. Vis. Pattern Recognit.*, pp. 3640–3649, 2016.

[49] J. Schlemper *et al.*, "Attention gated networks: learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, 2019.

[50] J. Schlemper *et al.*, "Learn to pay attention," *ICLR*, 2018.

[51] O. Oktay *et al.*, "Attention U-net: Learning where to look for the pancreas," in *Proc. Med. Imaging with Deep Learning*, 2018.

[52] O. Çiçek *et al.*, "3D Unet: learning dense volumetric segmentation from sparse annotation," *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, pp. 424–432, 2016.

[53] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," *IEEE 3D Vision*, pp. 565–571, 2016.

[54] R. Gu *et al.*, "CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation," *IEEE Trans. Med. Imaging*, vol. 40, no. 2, pp. 699–711, 2021.

[55] ,C. Guo *et al.*, "Saunet: Spatial attention u-net for retinal vessel segmentation," *Int. Conf. Pattern Recognit.*, pp. 1236–1242, arXiv preprint arXiv:2004.03696, 2020.

[56] , F. Wang *et al.*, "Residual attention network for image classification," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* pp. 6450–6458, 2017.

[57] Z. Cheng *et al.*, "Contour-aware semantic segmentation network with spatial attention mechanism for medical image," *Vis. Comput.*, vol. 38, no. 3, pp. 749–762, 2022.

[58] Q. Huang *et al.*, "Dual-Term Loss Function For Shape-Aware Medical Image Segmentation," in *Proc. IEEE Int. Symp. Biomed. Imaging*, pp. 1798–1802, 2021.

[59] S. Park and M. Chung, "Cardiac segmentation on CT Images through shape-aware contour attentions," *Comput. Biol. Med.*, vol. 147, p. 105782, 2022.

[60] B. Mandelbrot *et al.*, "The Fractal Geometry of Nature," *W. H. Freeman and Company*, 1983.

[61] G. Castellano, L. Bonilha, L. Li, and F. Cendes, "Texture analysis of medical images," *Clin. Radiol.*, vol. 59, no. 12, pp. 1061–1069, 2004.

[62] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[63] K. He *et al.*, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *ICCV*, pp. 1026–1034, 2015.

[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, arXiv:1412.6980. [Online]. Available: https://arxiv.org/abs/1412.6980

[65] M. D. Zeiler *et al.*, "Adadelta: an adaptive learning rate method," *ICCV*, arXiv preprint arXiv:1212.5701, 2012.

[66] C. Tobon-Gomez *et al.*, "Benchmark for algorithms segmenting the left atrium from 3D CT and MRI datasets," *IEEE Trans. Med. Imaging*, vol. 34, no. 7, pp. 1460–1473, 2015.

[67] A. E. Kavur *et al.*, "CHAOS Challenge - Combined (CT-MR) healthy abdominal organ segmentation," *Med. Image Anal.*, vol. 69, pp. 101950, 2021.

[68] P. A. Yushkevich *et al.*, "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006.