Highlights

***MTFD-Net*: Left Atrium Segmentation in CT Images through Fractal Dimension Estimation**

Aziza Saber Jabdaragh, Marjan Firouznia, Karim Faez, Fariba Alikhani, Javad Alikhani Koupaei, Cigdem Gunduz-Demir

- It introduces an encoder-decoder network that quantifies texture complexity through self-similar patterns in an image.
- The model proposes to learn fractal dimension map estimation together with segmentation in a multi-task learning framework.
- The model led to improved performance for left atrium segmentation in computed tomography images.

Pattern Recognition Letters
journal homepage: www.elsevier.com

MTFD-Net: Left Atrium Segmentation in CT Images through Fractal Dimension Estimation

Aziza Saber Jabdaragh[a], Marjan Firouznia[b], Karim Faez[b], Fariba Alikhani[c], Javad Alikhani Koupaei[d], Cigdem Gunduz-Demir[e,f,g,**]

[a]*Department of Computer Engineering, Bilkent University, Ankara, Turkey*
[b]*Department of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran*
[c]*Department of Radiology, Isfahan University of Medical Sciences, Isfahan, Iran*
[d]*Department of Mathematics, Payame Noor University, Tehran, Iran*
[e]*Department of Computer Engineering, Koc University, Istanbul, Turkey*
[f]*School of Medicine, Koc University, Istanbul, Turkey*
[g]*KUIS AI Center, Koc University, Istanbul, Turkey*

Article history:

Fractal dimension, multi-task learning, dense prediction networks, computed tomography, segmentation

ABSTRACT

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 paper 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 (LA) segmentation in computed tomography (CT) 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 LA segmentations compared to its counterparts.

© 2023 Elsevier Ltd. All rights reserved.

1. Introduction

Encoder-decoder networks have shown great promise for a wide range of medical image segmentation problems, including the segmentation of computed tomography (CT) images. Nevertheless, it is still challenging for these networks to segment a region of interest (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 better segment the RoIs, previous 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 [9]. Another group exploited this information in the form of designing a multi-task network. They typically defined learning contour [6] and shape [3], [15] 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

**Corresponding author
e-mail: aziza.saber@bilkent.edu.tr (Aziza Saber Jabdaragh), marjan.abdechiri@aut.ac.ir (Marjan Firouznia), kfaez@aut.ac.ir (Karim Faez), fariba.alikhani@med.mui.ac.ir (Fariba Alikhani), J.alikhani48@pnu.ac.ir (Javad Alikhani Koupaei), cgunduz@ku.edu.tr (Cigdem Gunduz-Demir)

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 may not follow the smoothness constraint in the Euclidean geometry, especially when the RoI is a part of a soft organ and contains intralesional heterogeneity. 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 [18].

In response to this issue, this paper presents a new dense prediction neural network design, which we name *MTFD-Net*. In this design, we propose to calculate 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 left atrium (LA) segmentation in CT images. The main contributions of this paper 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 [1]. 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 [1], [16], these studies did not utilize fractal textures in a dense prediction 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 at the same time. 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 [4]. Second, the shared feature representation, learned at the various layers of the shared encoder path, should keep necessary context and texture information to realize LA segmentation and FD map estimation, respectively. This enforces the network to more effectively incorporate additional texture information into the segmentation process. Although there exist studies that developed multi-task network architectures for various applications [3], [6], [8], [15], [27], none of them defined their auxiliary task using the fractal geometry.

2. Related Work

Network Architectures: The U-Net was one of the first successful encoder-decoder networks for medical image segmentation [24]. Although it achieved promising results for various problems, it may not accurately segment RoIs with texture variations and may fail to precisely delineate irregular RoI boundaries. Thus, many studies proposed U-Net variants that employed contextual, shape, and contour information to estimate better RoIs. One group used the attention mechanism. In [20], attention gates were added to a U-Net backbone to preserve the spatial contextual information for pancreas segmentation in CT images. In [9], an attention mechanism was added to utilize the shape information for multiple organ segmentation in abdominal CT scans. In [10], multiple modules were defined to make the network focus on foreground regions, modulate channel-wise feature responses, and emphasize salient feature maps at multiple scales. Other studies also used multiple modules to extract complementary features, which were then aggregated for the final segmentation [17], [22].

The other group employed contextual, shape, and contour information in the form of designing a multi-task network. They typically defined contour estimation as an auxiliary task and learned it with the main task of RoI segmentation [6], [23], [30]. In [7], one semantic-branch was used to capture semantic information using a spatial attention module and one detail-branch to focus on contour information in its shallow layers. Likewise, in [27], contour information was used in the form of a contextual network that utilized pyramid edge detection, multi-task, and interactive attention modules to propagate salient context information. It was proposed to define a shape-aware loss in multi-task networks [15] and to employ distance transforms as auxiliary tasks for developing shape-aware modules in these networks [8], [21]. Nevertheless, none of these previous studies used the fractal geometry to construct their attention mechanisms or to define their auxiliary tasks.

Fractal Dimension: Fractals are the structures that exhibit similar patterns at different scales. This characteristic is known as self-similarity, complexity of which is typically quantified by the fractal dimension [18]. Many natural surfaces exhibit self-similarity characteristics. Rough surfaces typically have more irregular textures characterized with higher fractal dimensions [19]. For instance, it was shown that aggressive tumors had higher fractal dimensions than non-aggressive ones [1]. There are different techniques, including the box-counting method, to transform pixel intensities into their fractal dimensions [2]. Since its first introduction by [18], the fractal features have been widely used in many applications including medical image analysis [16]. For instance, the fractal dimension was used to differentiate internal and peripheral textures of bronchogenic and bronchioloalveolar cell carcinomas [13] and to characterize lung tissues [1] in CT images. On the other hand, these previous studies did not make use of any fractal features to design a multi-task network for a dense prediction task in CT images.

3. Methodology

The *MTFD-Net* model relies on 1) quantifying the texture complexity and variation through calculating a map of fractal dimensions and 2) designing and training a multi-task network that utilizes the FD map to provide auxiliary supervision to the main task of segmentation. The details are given below.



(a) (b) (c)

Fig. 1. (a) An example gray-scale CT image. (b) Ground truth of its LA region. (c) Its FD map generated using the differential box-counting method [1].

3.1. Fractal Dimension Map Generation

The fractal dimension provides an effective means of quantifying rough surfaces in medical images [5], especially when a surface belonging to an organ or a biological structure exhibits similar patterns at different scales. In this work, 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 [1], which used the differential box-counting method that works on gray-level intensities [25].

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 follow. First, for the scaling factors $2 \leq r \leq R$, the boxes $N_r(x, y)$ are calculated as given in Eqn. 1, 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 $\log(N_r(x, y))$.

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

where

$$M_r(x, y) = \max_{(u,v) \in w_r(x,y)} I(u, v) \tag{2}$$

$$m_r(x, y) = \min_{(u,v) \in w_r(x,y)} I(u, v). \tag{3}$$

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 empirically chose $R = 7$.

The FD map generated for an exemplary CT image together with its ground truth map are shown in Fig. 1. As shown in this figure, the FD map provides a texture representation complementary to the ground truth map. Additionally, the fractal dimensions of boundaries between different types of regions exhibit high contrast differences. Learning this FD map from the original gray-level CT image provides auxiliary supervision to the network, towards betterment of LA segmentation. The details of this learning will be discussed in the next subsection.

3.2. Multi-Task Network Architecture

MTFD-Net uses a multi-task architecture to learn LA segmentation and FD estimation at the same time. This architecture has one shared encoder and two separate decoders, one for LA segmentation and the other for FD estimation (Fig. 2). The decoder paths decode their output maps from the same feature representations learned at different layers of the shared encoder. Thus, the encoder needs to keep 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 effective to mitigate the overfitting problem, which would more likely happen while learning LA segmentation with a single-task network [4].

The encoder path includes four blocks of two convolutions and one max pooling. The convolution layers use a 3×3 filter and followed by the rectified linear unit (ReLu) activation function. The dropout layer with a dropout factor of 0.2 is added to prevent overfitting. The max pooling layer uses a 2×2 filter to reduce the spatial dimension by half at the end of each block. The bottleneck block has the same two convolution layers without max pooling. 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 fed to the convolution operators. The number of feature maps are halved at the end of each decoder block. At the end, the softmax 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 Fig. 2.

3.3. Implementation Details

MTFD-Net was trained to minimize a joint loss function. This 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. Their contributions to the joint loss were the same. The pixel contributions in the weighted categorical cross entropy were selected inversely proportional to the class frequencies. The network hyperparameters used in training are listed in Table 1. Note that here we set the batch size as 1 considering the size of our network architecture and our available GPUs; selecting larger batch sizes

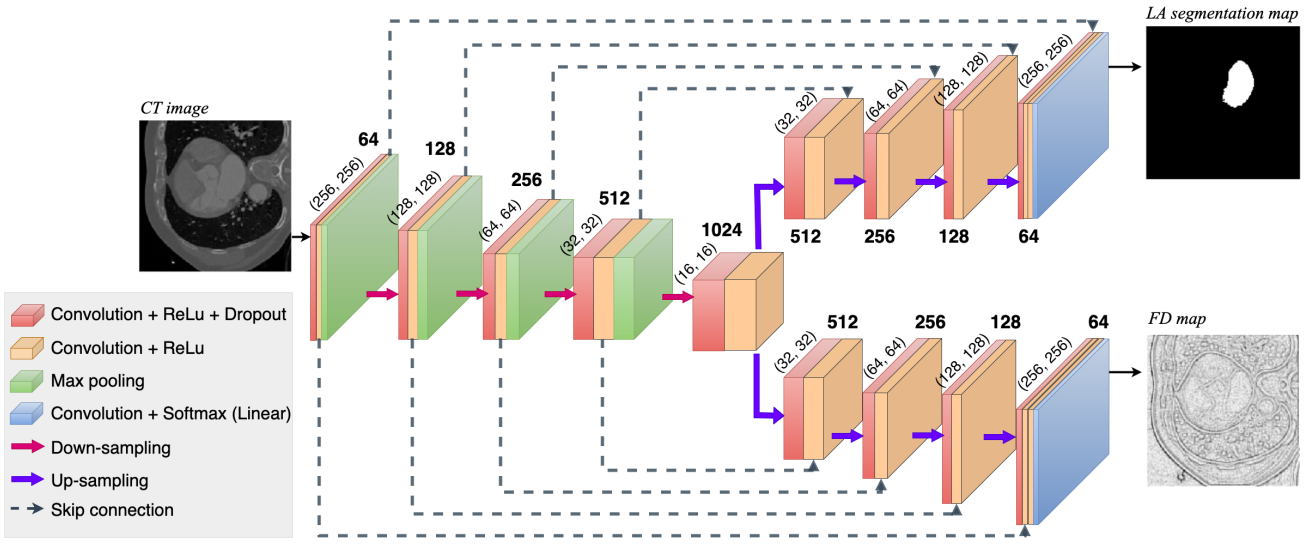


Fig. 2. Multi-task architecture of *MTFD-Net*. Each box and arrow corresponds to an operation distinguishable by its color. The numbers (h, w) on the left of a block denote the block's input height and width, respectively. The number of feature maps used by a block is indicated at its top (or bottom) in bold.

Table 1. Network hyperparameters used in training.

Maximum iteration	600 epochs
Patience for early stopping	100 epochs
Optimizer	Adam ($\beta_1 = 0.5, \beta_2 = 0.999$)
Learning rate	1e-5
Batch size	1

will increase the required GPU memory. All models were implemented in Python under the open-source deep learning libraries, Tensorflow and Keras. The network implementation is available at mysite.ku.edu.tr/cgunduz/downloads/mtfd-net/.

The implemented *MTFD-Net* model has a second decoder for FD map estimation, which made the network's training time longer than that of its single-task counterpart. Additionally, FD maps were calculated as the target output of this second decoder. The map calculation of a single image took 48 seconds on the average on an Intel i7-5557U 3.10 GHz CPU. The multi-task network training took 18.6 hours on the average on a Tesla T4 GPU whereas the single-task network took 8.1 hours. However, to segment an unseen test image, only the first decoder's estimation is used, and there is no need to run the second decoder. Moreover, since the FD map is used as the target output of the second decoder, but not as an input, its calculation is only necessary for training. These made the testing time of the multi-task network and its single-task counterpart comparable. For a single image, the testing time of our network was 38.0 milliseconds on the average on an Tesla T4 GPU.

4. Experiments

4.1. Dataset

We tested the proposed model on an left atrium (LA) dataset provided by a retrospective study involving a cohort of subjects with atrial fibrillation. The cohort was obtained by University Hospitals from Iran in 2017-2019. The dataset consisted of 2560 CT images of 20 subjects (128 images from the CT scan of each subject). Ground truth segmentations were obtained following an approach used by [26]: initial segmentation maps were created using a 3D region growing algorithm [26] and manual enhancements on these initial maps were performed by a radiologist with six years experience. We used three-fold cross-validation in the experiments. For that, the 20 subjects were randomly split into three folds. Then, for each fold, the CT images of the subjects in that fold were used as the test set and the rest as the training set. The training set was further split into training images, on which the network weights were learned by backpropagation, and validation images, which were used for early stopping.

4.2. Evaluation

LA segmentations were evaluated visually and quantitatively on the images of the test folds. 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 precision, recall, and Dice index were calculated. These metrics were to assess how accurately a model segmented the foreground pixels. In Table 2, these metrics were reported in two different ways. First, they were calculated for each 2D CT image separately and averaged over all test set images that contain at least one RoI in their ground truths (*averaged metrics*). Then, all TP, FP, and FN pixels were accumulated over all test set images, regardless of whether they contain an RoI in the ground truth or not, and then precision, recall, and Dice score were calculated on those accumulated numbers (*accumulated metrics*). It is also worth noting that CT images equally contributed to the averaged metrics, regardless of the size of their LA regions, whereas they contributed to the accumulated metrics proportional to their LA regions' sizes.

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 sym-

Table 2. Cross-validation test set results obtained by the proposed *MTFD-Net* models and the comparison algorithms. These are the average scores and standard deviations across 15 runs (three folds and five runs for each fold). Significantly best metrics ($p < 0.05$) are indicated with blue font color. Note that for a selected performance metric, there is no statistically significant difference between the values that are all shown in blue.

	Averaged			Accumulated				
	Precision	Recall	Dice index	Precision	Recall	Dice index	ASSD	MSSD
MTFD-Net	84.4 ± 1.5	83.6 ± 2.7	**80.8 ± 1.6**	89.9 ± 1.0	91.4 ± 1.3	**90.7 ± 0.6**	**2.6 ± 0.6**	**55.2 ± 16.0**
MTFD-Net-2	**87.5 ± 1.7**	82.2 ± 2.2	**81.3 ± 1.1**	**91.7 ± 1.6**	90.8 ± 1.0	**91.2 ± 0.9**	**2.7 ± 0.8**	**59.6 ± 19.5**
Single-task UNet [24]	79.1 ± 2.9	84.9 ± 4.3	79.0 ± 2.8	85.1 ± 4.0	90.8 ± 2.3	87.7 ± 1.9	4.0 ± 1.7	85.0 ± 18.3
Multi-task-shape [3]	80.2 ± 1.9	**86.4 ± 2.8**	**80.6 ± 2.3**	85.3 ± 2.7	**93.1 ± 1.1**	89.0 ± 1.2	4.5 ± 1.3	95.5 ± 14.3
Unet 3+ [11]	82.2 ± 3.0	74.1 ± 5.7	74.6 ± 2.9	86.9 ± 3.8	83.1 ± 4.0	84.8 ± 0.9	5.8 ± 2.9	79.2 ± 24.8
ShapePU [29]	**88.8 ± 3.4**	76.6 ± 6.1	76.9 ± 4.3	**91.1 ± 3.3**	88.1 ± 3.2	89.5 ± 0.8	**3.0 ± 1.1**	**52.3 ± 24.7**

metric surface distance (MSSD) were calculated as explained in [12]. These metrics were calculated on the boundary voxels S_B and G_B of the estimated segmentation and ground truth volumes of a 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 closest boundary voxel in the estimated volume was calculated. The ASSD and MSSD were the average and maximum of all these distances, respectively. These metrics were calculated on each CT volume and averaged over the subjects in the test folds. Better segmentations yield higher precision, recall, and Dice index, and lower ASSD and MSSD.

$$d(v, G_B) = \min_{u \in G_B} \|v - u\|^2 \tag{4}$$

$$d(u, S_B) = \min_{v \in S_B} \|u - v\|^2 \tag{5}$$

$$ASSD = \frac{\sum_{v \in S_B} d(v, G_B) + \sum_{u \in G_B} d(u, S_B)}{|G_B| + |S_B|} \tag{6}$$

$$MSSD = \max\left(\max_{v \in S_B} d(v, G_B), \max_{u \in G_B} d(u, S_B)\right) \tag{7}$$

4.3. Comparisons

We used five algorithms for comparison and ablation studies. The first one was the baseline algorithm that used a standard UNet architecture [24]. It was a single-task network with one encoder and one decoder that took a CT image as its input and predicted an LA segmentation map. All operators in these encoder and decoder were the same with those specified in Fig. 2. The second algorithm (*multi-task-shape*) also used a multi-task network, with exactly the same encoder and decoder architectures shown in Fig. 2. On the contrary, instead of employing a fractal dimension map, it defined another auxiliary task to employ shape information in the network design. For that, it extracted Fourier descriptors on the ground truth maps, as explained in [3], and defined the learning of these Fourier descriptors as its auxiliary task. As opposed to *MTFD-Net*, which used the fractal geometry to model the complexity and variation in the texture of an RoI, the *multi-task-shape* algorithm modeled the shape of the RoI using the Fourier descriptors defined in the Euclidean geometry.

These two comparison algorithms were also considered as part of ablation studies. We used the first one to understand the effectiveness of using an additional task in our network training, and the second one to compare the effects of using different forms of additional information modeled by the fractal and Euclidean geometries. As a third ablation study, we implemented an alternative differential box counting method to calculate FD maps and used these maps to define the auxiliary task in our network design. For that, we revised the computation-quantization mechanism of [1] by incorporating the approach followed by [14]. Both of these methods used non-overlapping partitioning, but the method of [1] suffered from over-counting the boxes along the directions. In contrast, the alternative method addressed this issue by modifying the non-overlapping partitioning mechanism to include a continuous surface on the boundary of neighboring grids. We considered this alternative, which we named *MTFD-Net*-2, to demonstrate that one may use a different FD calculation method, and still obtain comparable results. All these three algorithms trained their networks using the same setup with *MTFD-Net*.

The fourth algorithm, the UNet 3+ model, was a modified UNet architecture that was trained to minimize a hybrid loss function [11]. This modified version defined extra skip connections from encoder to decoder blocks as well as intra-connections among the decoder blocks to make use of multi-scale features in decoding a segmentation map. Additionally, each of the decoder blocks was upsampled and supervised by the ground truth. Its hybrid loss function combined multi-scale structural similarity index loss with focal and intersection-over-union losses. The last algorithm was the ShapePU model that presented a shape-contrained positive-unlabeled learning framework [29]. It was a UNet-based model that also sought supervision from unlabeled pixels by adopting EM estimation. It applied cutout operations on training images and defined a shape-consistency loss to penalize inconsistent segmentation results between the original training images and their cutout versions. Even though it could work with scribble annotations, in our experiments, we turned off this option and provided fully annotated maps to this model for fair comparison.

4.4. Results

The quantitative results obtained on the test folds were presented in Table 2. Note that for all models, we performed three-fold cross-validation, and for each fold, we trained the network five times. This table indicated the average scores and standard

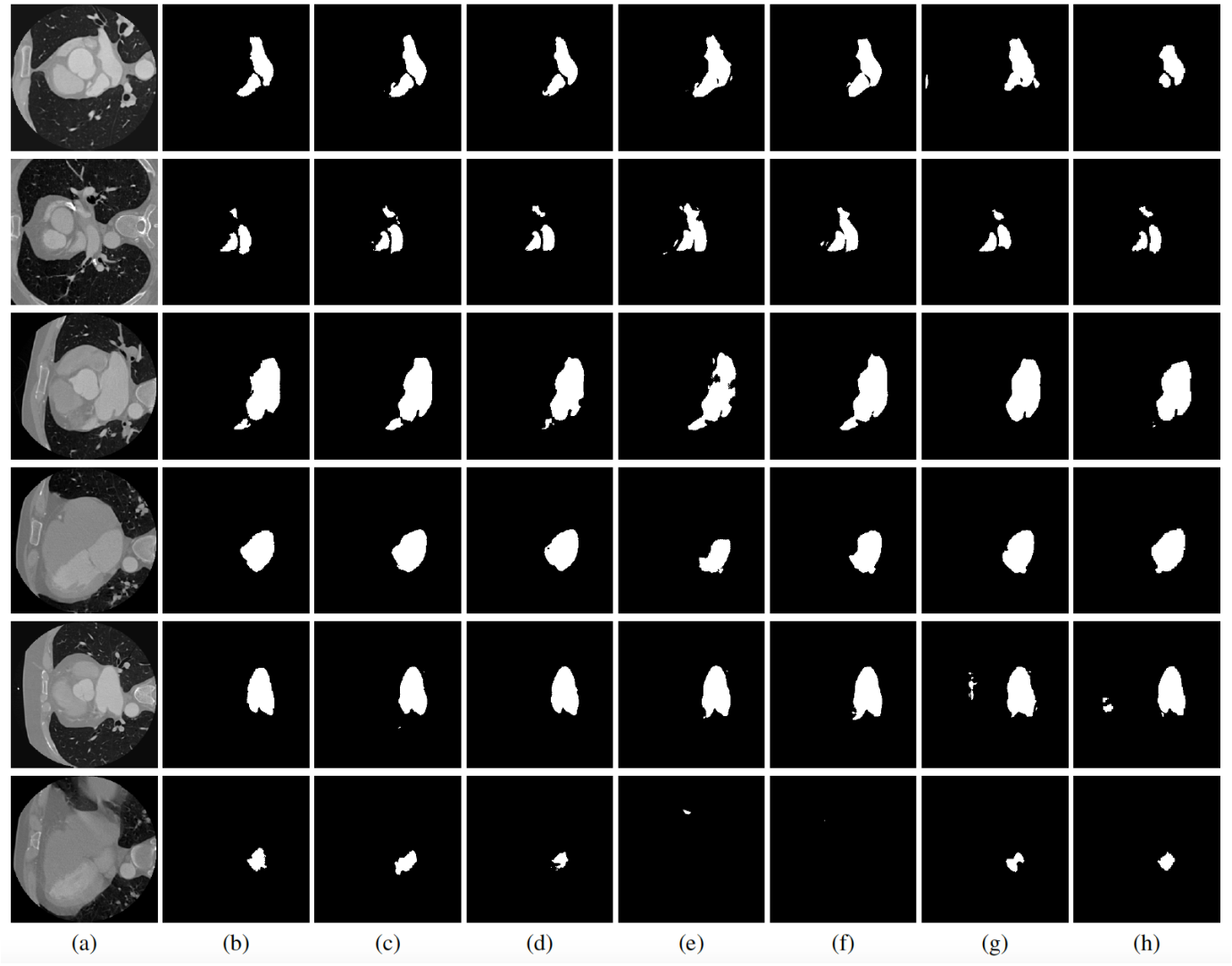


Fig. 3. (a) Example test set images. (b) Ground truths. Results of (c) the proposed *MTFD-Net* that uses the differential box-counting method proposed by [1], (d) the proposed *MTFD-Net*-2 that uses the differential box-counting method proposed by [14], (e) the single-task UNet [24], (f) the multi-task-shape [3], (g) the UNet 3+ [11], and (h) the ShapePU [29] algorithms.

deviations across these 15 runs. We also applied the paired-sample t-test on the results to understand whether the differences were statistically significant.

This table revealed that the proposed *MTFD-Net* model and its alternative *MTFD-Net*-2 led to high pixel-level scores and low distance-based metrics. Comparing them with the single-task UNet algorithm and the UNet 3+ model, which had extra skip connections and a hybrid loss function, the 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 [4]. The differences in all metrics were statistically significant with $p < 0.05$.

The accumulated Dice index and the distance metrics of the proposed models were statistically significantly better than those of the *multi-task-shape* comparison algorithm. The averaged Dice index scores, which were computed only on the images with at least one RoI in their ground truths, were similar ($p < 0.05$). This comparison 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.

Comparing with the *ShapePU* algorithm, the proposed *MTFD-Net* model and its alternative *MTFD-Net*-2 yielded statistically better Dice index scores; the distance metrics (ASSD and MSSD) were statistically similar ($p < 0.05$). This might be attributed to the following: *ShapePU* found the LA regions more compact, which decreased the distance between an estimated voxel and the closest boundary voxel and vice versa. This, in turn, decreased the calculated ASSD and MSSD. On the other hand, finding the LA regions more compact than the ground truths resulted in lower Dice index scores. This was also observed in the precision and recall metrics; compact regions decreased the number of FPs (higher precision) while increasing the number of FNs (lower recall). Note that we did not directly use the precision and recall metrics for comparing the performance of two algorithms since one algorithm may give

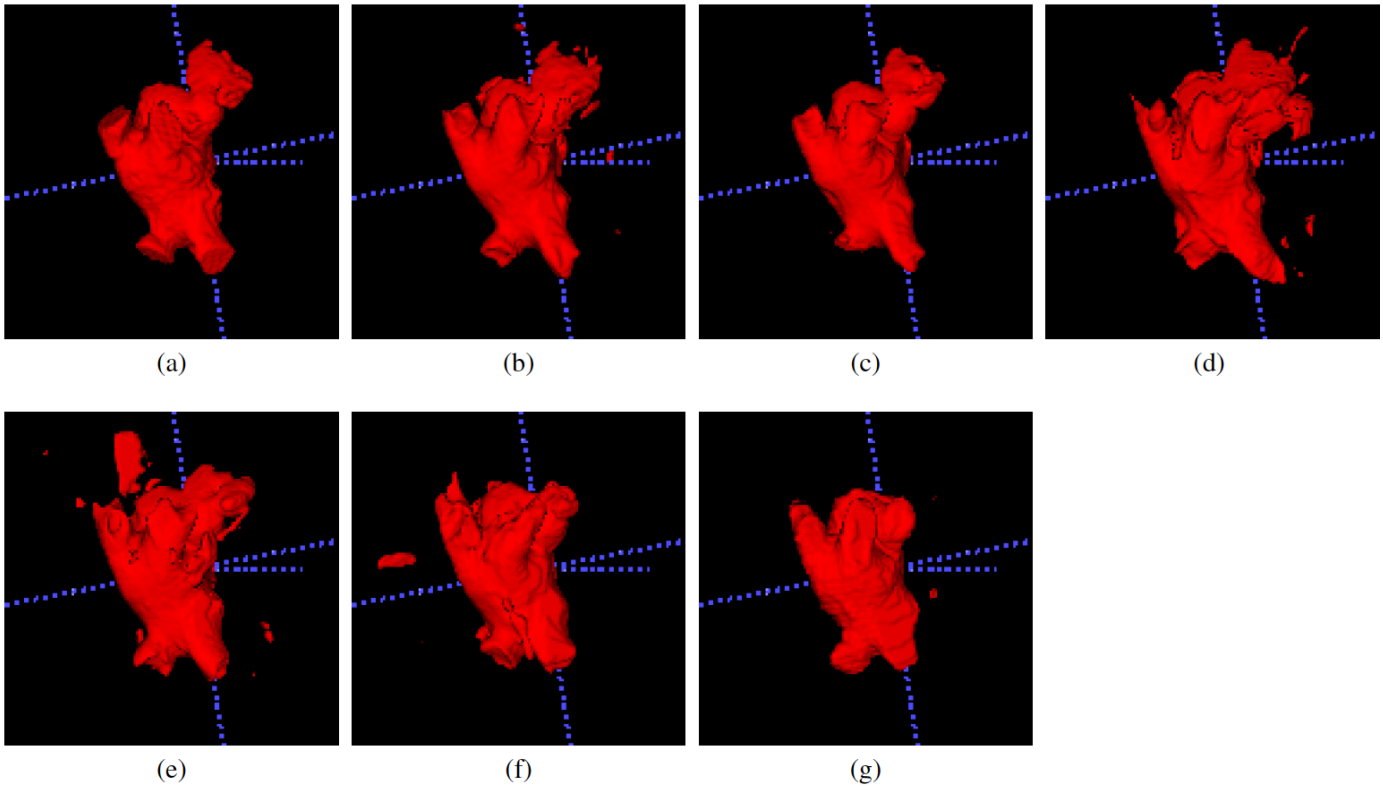


Fig. 4. For an exemplary test set subject, LA construction (a) from the 3D volume of the ground truth maps, and from the 3D volume of the segmentation maps generated by (b) the proposed *MTFD-Net* that uses the differential box-counting method proposed by [1], (c) the proposed *MTFD-Net*-2 that uses the differential box-counting method proposed by [14], (d) the single-task UNet [24], (e) the multi-task-shape [3], (f) the UNet 3+ [11], and (g) the ShapePU [29] algorithms.

better recall but worse precision or vice versa. It is very well known that there is a trade-off between precision and recall, which is quantified with the Dice index score. Nevertheless, the proposed models achieved high Dice index scores and low distance metrics at the same time.

Lastly, we compared the *MTFD-Net* and *MTFD-Net*-2 models, which used different differential box counting methods to calculate their FD maps. The *MTFD-Net* led to slightly better ASSD and MSSD metrics but slightly worse Dice index scores. However, the paired-sample t-test showed that there was no statistically significant difference between these scores ($p < 0.05$). This experiment revealed that our network design allowed to use alternative differential box counting methods for FD map calculation. It also indicated the effectiveness of using fractal geometry in an auxiliary task definition for LA segmentation.

These quantitative results were also consistent with visual results shown in Fig. 3. As seen in the first three rows of this figure, the *MTFD-Net* and *MTFD-Net*-2 models more successfully delineated the boundaries of adjacent LA regions. Additionally, when there is only one LA region in the image, they led to better-shaped estimations (the fourth and fifth rows), even though the *multi-task-shape* comparison algorithm explicitly modeled the shape information of an RoI. As seen in the last row, all algorithms might fail for some images. Even on such an image, we observed that the *MTFD-Net* models still gave similar or better segmentations, compared with the other algorithms.

Additionally, Fig. 4 depicted the 3D LA construction for an exemplary test set subject. These were constructed by ITK-snap, an open-source, multi-platform medical image visualization and segmentation tool [28], on the ground truths as well as the estimated segmentation maps. This figure showed that the LA region estimations of the proposed *MTFD-Net* model and its alternative *MTFD-Net*-2 led to more accurate constructions than the comparison algorithms.

5. Conclusions

This paper presented a new multi-task network design, which we named *MTFD-Net*, for LA segmentation in CT images. The *MTFD-Net* model relied on benefiting the fractal geometry to provide auxiliary supervision to the main task, towards the betterment of LA segmentation. To this end, it transformed a CT image into a map of fractal dimensions to quantify texture complexity through self-similar patterns in the image and incorporated its learning into a multi-task network design. This was the first proposal of a dense prediction network that employed the fractal geometry to define an auxiliary task and learned it in parallel with the main task of segmentation in a multi-task learning framework.

We tested our *MTFD-Net* model for LA segmentation on 2560 CT images of 20 subjects. Our experiments revealed that transforming an image into a map of fractal dimensions and

concurrently learning it with the main segmentation task led to improved performances compared to its counterparts.

One limitation of *MTFD-Net* was to model the texture of an RoI but not its shape. It could be possible to design multi-task networks that would utilize the shape-related auxiliary tasks. This is one future research direction of this work. The *MTFD-Net* model used LA segmentation as a showcase application due to its importance in analyzing various heart-related diseases such as atrial fibrillation. Another future research direction is to explore the use of *MTFD-Net* for segmenting different parts of the heart in CT images.

Acknowledgment

This work was supported by the Scientific and Technological Research Council of Turkey, project no: TÜBİTAK 220N354 and International Academic Cooperation Directorate University of Tabriz and Ministry of Science, Research and Technology of Iran, Project no: rRTU-2-1402, 1400-05-01.

References

[1] O. S. Al-Kadi and D. Watson. Texture analysis of aggressive and nonaggressive lung tumor ce ct images. *IEEE Trans. Biomed. Eng.*, 55(7):1822–1830, 2008.

[2] A. Balghonaim and J. Keller. A maximum likelihood estimate for two-variable fractal surface. *IEEE Trans. Image Process.*, 7(12):1746–1753, 1998.

[3] S. Cansiz, C. Kesim, S. N. Bektas, Z. Kulali, M. Hasanreisoglu, and C. Gunduz-Demir. Fouriernet: Shape-preserving network for henle's fiber layer segmentation in optical coherence tomography images. *IEEE J. Biomed. Health Inform.*, 27:1036–1047, 2023.

[4] R. Caruana. Multitask learning. *Mach. Learn.*, 28:41–75, 1997.

[5] G. Castellano, L. Bonilha, L. Li, and F. Cendes. Texture analysis of medical images. *Clin. Radiol.*, 59(12):1061–1069, 2004.

[6] 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.*, pages 2487–2496, 2016.

[7] Z. Cheng, A. Qu, and X. He. Contour-aware semantic segmentation network with spatial attention mechanism for medical image. *Vis. Comput.*, 38(3):749–762, 2022.

[8] 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.*, 113:102023, 2021.

[9] E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson, and D. C. Barratt. Automatic multi-organ segmentation on abdominal ct with dense v-networks. *IEEE Trans. Med. Imaging*, 37(8):1822–1834, 2018.

[10] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang. Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE Trans. Med. Imaging*, 40(2):699–711, 2011.

[11] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 1055–1059, 2020.

[12] A.E. Kavur, N.S. Gezer, M. Baris, S. Aslan, P.-H. Conze, V. Groza, D.D. Pham, S. Chatterjee, P. Ernst, S. Ozkan, B. Baydar, D. Lachinov, S. Han, J. Pauli, F. Isensee, M. Perkonigg, R. Sathish, R. Rajan, D. Sheet, G. Dovletov, O. Speck, A. Nurnberger, K.H. Maier-Hein, G.B. Akar, G. Unal, O. Dicle, and M.A. Selver. Chaos challenge - combined (ct-mr) healthy abdominal organ segmentation. *Med. Image Anal.*, 69:101950, 2021.

[13] S. Kido, K. Kuriyama, M. Higashiyama, T. Kasugai, and C. Kuroda. 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.*, 27(1):56–61, 2003.

[14] K. Lai, C. Li, T. He, L. Chen, K. Yu, and W. Zhou. Study on an improved differential box-counting approach for gray-level variation of images. In *10th International Conference on Sensing Technology*, pages 1–6, 2016.

[15] H. Li, X. Liu, S. Boumaraf, X. Gong, D. Liao, and X. Ma. Deep distance map regression network with shape-aware loss for imbalanced medical image segmentation. In *Machine Learning in Medical Imaging*, pages 231–240, 2020.

[16] R. Lopes and N. Betrouni. Fractal and multifractal analysis: A review. *Med. Image Anal.*, 13(4):634–649, 2009.

[17] K.-Y. Lung, C.-R. Chang, S.-E. Weng, H.-S. Lin, H.-H. Shuai, and W.-H. Cheng. Rosnet: Robust one-stage network for ct lesion detection. *Pattern Recognit. Lett.*, 144:82–88, 2021.

[18] B. B. Mandelbrot. *Fractal Geometry of Nature*. W. H. Freeman and Co., 1982.

[19] W. H. Nailon. Texture analysis methods for medical image characterisation. *Biomedical Imaging*, 75, 2010.

[20] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert. Attention u-net: Learning where to look for the pancreas. In *Proc. Med. Imaging with Deep Learning*, 2018.

[21] S. Park and M. Chung. Cardiac segmentation on ct images through shape-aware contour attentions. *Comput. Biol. Med.*, 147:105782, 2022.

[22] L. Qu, M. Wang, K. Guo, W. Wan, Y. Liu, J. Tang, J. Wu, and P. Duan. Biomedical image segmentation based on full-resolution network. *Pattern Recognit. Lett.*, 155:232–238, 2022.

[23] J. Ren, H. Sun, H. Zhao, H. Gao, C. Maclellan, S. Zhao, and X. Luo. Effective extraction of ventricles and myocardium objects from cardiac magnetic resonance images with a multi-task learning u-net. *Pattern Recognit. Lett.*, 155:165–170, 2022.

[24] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, page 234–241, 2015.

[25] N. Sarkar and B. B. Chaudhuri. An efficient differential box-counting approach to compute fractal dimension of image. *IEEE Trans. Syst. Man Cybern.*, 24(1):115–120, 1994.

[26] C. Tobon-Gomez, A.J. Geers, J. Peters, J. Weese, K. Pinto, R. Karim, M. Ammar, A. Daoudi, J. Margeta, Z. Sandoval, B. Stender, Y. Zheng, M. A. Zuluaga, J. Betancur, N. Ayache, M.A. Chikh, J.-L. Dillenseger, M.B. Kelm, S. Mahmoudi, S. Ourselin, A. Schlaefer, T. Schaeffter, R. Razavi, and K.S. Rhode. Benchmark for algorithms segmenting the left atrium from 3d ct and mri datasets. *IEEE Trans. Med. Imaging*, 34(7):1460–1473, 2015.

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

[28] P.A. Yushkevich, J. Piven, H.C. Hazlett, R.G. Smith, S. Ho, J.C. Gee, and G. Gerig. User-guided 3d active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage*, 31(3):1116–1128, 2006.

[29] K. Zhang and X. Zhuang. Shapepu: A new pu learning framework regularized by global consistency for scribble supervised cardiac segmentation. In *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, 2022.

[30] Q.-L. Zhang and Y.-B. Yang. A boundary-preserving conditional convolution network for instance segmentation. *Pattern Recognit. Lett.*, 163:1–9, 2022.