



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

공학석사학위논문

Contour and Distance Transform Guided  
Shape-aware Attention Network for Cardiac  
Segmentation

심장 분할을 위한 윤곽선 및 거리 변환 기반 모양 인식  
어텐션 네트워크

2021년 2월

서울대학교대학원

컴퓨터공학부

박 상 욱



Contour and Distance Transform Guided Shape-aware  
Attention Network for Cardiac Segmentation

심장 분할을 위한 윤곽선 및 거리 변환 기반 모양 인식  
어텐션 네트워크

지도교수 신 영 길

이 논문을 공학석사 학위논문으로 제출함

2020 년 11 월

서울대학교 대학원

컴퓨터공학부

박 상 옥

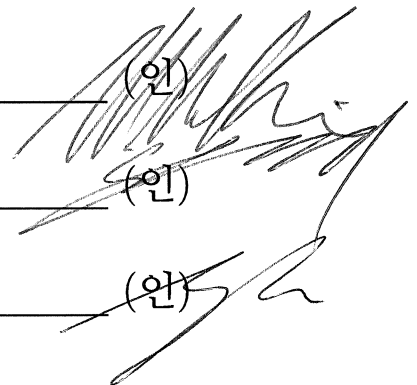
박 상 옥의 공학석사 학위论문을 인준함

2020 년 12 월

위 원 장 \_\_\_\_\_ 김 명 수 (인)

부위원장 \_\_\_\_\_ 신 영 길 (인)

위 원 \_\_\_\_\_ 이 영 기 (인)





# Abstract

Cardiac image segmentation is an important task in the development of clinical cardiac applications. Many recent studies came up with deep learning models, mostly composed of convolutional neural networks(CNN), and showed significant outcomes on the segmentation of target organs in medical images. Unlike other major biological structures such as lungs and liver, the cardiac organ consists of multiple substructures. Those cardiac substructures are intimately adjacent to each other, which means that the segmentation network should concentrate on the boundaries of the substructures. In this paper, to increase the performance of cardiac image segmentation, we introduce a novel model to learn shape-aware and boundary-aware features using the distance transformation and the contour image of the labeled data. We present a shape-aware attention module that can guide a model to focus on edges between structures. we also propose the regularization for refining the contour probabilistic map. The experimental results show that the proposed network produces more accurate results compared to state-of-the-art networks by obtaining 4.97% more in terms of dice similarity coefficient(DSC) score. We used 20 CT cardiac images for training and validation, and 40 CT cardiac images for the test. Moreover, our segmentation results describe that learning precise contour and distance transform features can help improve model performance. The ablations studies are presented to emphasize the importance of our proposed shape-aware attention mechanism.

**Keywords:** deep neural network, shape-aware attention, biomedical image segmentation, cardiac segmentation

**Student Number:** 2019-23191

# Contents

<b>Abstract</b>	<b>i</b>
<b>Chapter 1 Introduction</b>	<b>1</b>
<b>Chapter 2 Backgrounds</b>	<b>5</b>
2.1 Deep Neural Networks for Image Segmentation . . . . .	5
2.1.1 Overview . . . . .	5
2.1.2 Convolutional Layer . . . . .	6
2.1.3 Residual Connection . . . . .	8
2.2 Visual Attention Mechanism . . . . .	9
<b>Chapter 3 Related Works</b>	<b>11</b>
3.1 CNNs for Biomedical Image Segmentation . . . . .	11
3.2 Distance Transformation for CNN Image Segmentation . . . . .	14
3.2.1 Distance Transformation(DT) . . . . .	14
3.2.2 DT-based Image Segmentation . . . . .	15
<b>Chapter 4 Methodology</b>	<b>18</b>
4.1 Overview . . . . .	18
4.2 CDA-Net Architecture . . . . .	18

4.2.1	DT Regression for Image Segmentation . . . . .	19
4.2.2	Shape-aware Attention . . . . .	21
4.2.3	Regularization with the Penalty Term to Refine Feature . . . . .	22
4.3	Overall Loss Function . . . . .	23
4.4	Data Preparation . . . . .	25
<b>Chapter 5</b>	<b>Experimental Results</b>	<b>28</b>
5.1	Dataset . . . . .	28
5.2	Evaluation Metrics . . . . .	29
5.3	Experiments Detail . . . . .	30
5.4	Comparison with State-of-the-Art . . . . .	30
5.5	Ablation Study . . . . .	31
<b>Chapter 6</b>	<b>Conclusion</b>	<b>37</b>
	<b>Bibliography</b>	<b>39</b>
	<b>초록</b>	<b>45</b>



# List of Figures

Figure 1.1	Heart structure . . . . .	2
Figure 1.2	Example data . . . . .	3
Figure 1.3	Cardiac CT image . . . . .	3
Figure 2.1	Operation of convolutional layer . . . . .	7
Figure 2.2	Residual block . . . . .	8
Figure 2.3	Diagram of channel attention module and spatial attention module . . . . .	9
Figure 3.1	3-dimensional image segmentation . . . . .	12
Figure 3.2	U-Net architecture . . . . .	13
Figure 3.3	V-transition layer . . . . .	14
Figure 3.4	Comparison between the binary image and its distance transform . . . . .	15
Figure 3.5	Distance metrics for the distance transformation . . . . .	16
Figure 4.1	The architecture of Contour & Distance transform guided Attention NETWORK(CDA-Net) . . . . .	20
Figure 4.2	Shape-aware attention module . . . . .	22

Figure 4.3	The example case of penalty energy to refine contour and distance transform feature . . . . .	24
Figure 4.4	The example of input images . . . . .	27
Figure 5.1	The axial slices of segmentation results . . . . .	32
Figure 5.2	The model output visualization of cardiac structures . .	34
Figure 5.3	Visualization of the contour probabilistic map . . . . .	35

# List of Tables

Table 5.1	Dice similarity coefficient score of CDA-Net and the state-of-the-arts . . . . .	33
Table 5.2	Jaccard Index of CDA-Net and the state-of-the-arts . . .	33
Table 5.3	DSC of CDA-Net and its ablations . . . . .	36
Table 5.4	95% HD of CDA-Net and its ablations . . . . .	36
Table 5.5	ASSD of CDA-Net and its ablations . . . . .	36
Table 5.6	Sensitivity of CDA-Net and its ablations . . . . .	36
Table 5.7	Precision of CDA-Net and its ablations . . . . .	36

# Chapter 1

## Introduction

The heart is one of the most essential organs in the human body. Located in the center of the chest, it supplies blood flow throughout the body. This is why the heart is often described as the human body's engine. This vital organ operates constantly to supply blood flow, generally beating 100,000 times a day to deliver about 7,200 liters. This means that small abnormalities in the heart can lead the fatal consequences. Cardiovascular diseases(CVDs) is one of the major causes of death. [1] On average, someone dies of CVDs every 37 seconds and 2,353 death a day in the United States. [2] According to WHO, 17.9 million people die each year from CVDs.

However, it is estimated that up to 90% of CVDs can be preventable.[3] [4] To prevent CVDs, it is a fundamental process to diagnose the heart. With advances in computer technology and radiology, Computed Tomography(CT) and Magnetic Resonance Imaging(MRI) aids diagnosis with medical imaging. CT and MRI can be used to diagnose heart problems, which can greatly help prevent CVDs.

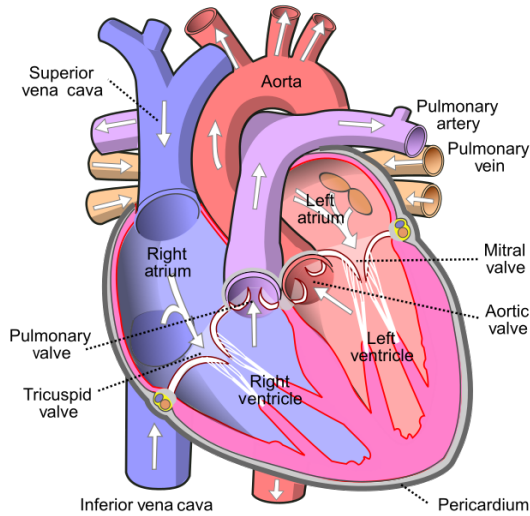


Figure 1.1: Heart structure[5]

In this paper, we aim at cardiac segmentation in CT images. It is a challenging problem since the heart is composed of several substructures. Fig. 1.2 shows that the human heart is a complex organ consisting of four chambers, blood vessels, and muscles. The target of this paper is to precisely segment substructures, including 4 chambers(LV: Left Ventricle, RV: Right Ventricle, LA: Left Atrium, RA: Right Atrium), 2 arteries(AA: Ascending Aorta, PA: Pulmonary Artery), and LV-myo: Myocardium of Left Ventricle. It means the major difficulty of segmentation is finding the boundary between adjacent organs. The fact that the heart is a huge muscle is another reason why it makes the problem fastidious. Due to the heartbeats every second, the contraction and relaxation of the heart allow the heart to be observed in various shapes in medical images. To address this problem, we proposed a novel attention mechanism. To focus on the shape and boundary of cardiac structures, we conjugate the features that imply the object’s contour and the distance transformation of labeled ground



Figure 1.2: Cardiac CT image, manual multi-organ segmentation image, and overlapped image

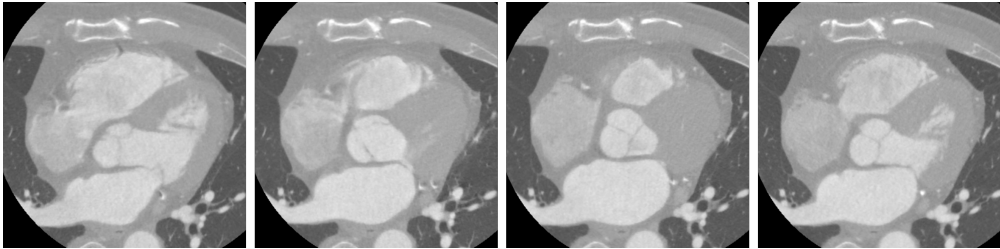


Figure 1.3: Cardiac CT image shows various shape in time domain sequence

truth images to the attention mechanism. Specifically, it makes the CNN model learn boundary-aware and shape-aware features. In other words, the proposed attention mechanism can produce exact segmentation results by reducing false-positive responses and finding accurate adjoining organ boundaries. We also introduce a simple idea to refine edge map and distance transform features. Besides, we designed the CNN model based on U-Net [6] encoder-decoder architecture and applied our proposed attention module.

The remainder of this paper is organized as follows. First, we introduce the backgrounds of deep learning including deep neural networks for image segmentation and an attention mechanism in chapter 2. Secondly, related researches are listed in chapter 3. Afterward, a novel attention mechanism using contour and

distance transformation is presented, then we evaluate our proposed method with the open dataset for whole heart segmentation in chapters 4 and 5. Subsequently, we compare the proposed method with other state-of-the-art models. In chapter 6, the conclusion is described.

# Chapter 2

## Backgrounds

### 2.1 Deep Neural Networks for Image Segmentation

#### 2.1.1 Overview

In the last 10 years, the performance of deep neural networks has skyrocketed. Likewise, deep neural networks revolutionized computer vision. Many traditional methods were replaced by deep learning techniques. Even the latest deep neural models surpass human ability. Deep learning has outperformed in many areas of computer vision, such as image classification, object detection, object localization, segmentation, style transfer, etc.

Convolutional Neural Network was proposed to apply a novel deep learning mechanism to visual data such as images and videos. *Yann LeCun et al.* [7] presented LeNet-5 in order to recognize digits from cropped text images. Thanks to soaring computing power, many CNN models have been suggested to handle bigger and larger datasets over the past decade.

AlexNet [8] turned up for image classification, GoogLeNet [9], VGGNet



[10], and Resnet [11] followed. Furthermore, in the object detection field, *Ross Girshick et al.* designed R-CNN [12] and Fast R-CNN[13], and *Shoqing Ren et al.* build them up to Faster R-CNN. [14] They select object candidates from image and crop to small boxes, then each box is classified as some object. Yolo [15] compressed the previous task to one step: finding a bounding box of object and classifying the box at the same time.

The concept of image classification can be extended to the image segmentation field. In the field of image segmentation, the interesting areas have generally annotated the foreground, and those that are not are marked as background. In this case, for instance, the segmentation problem can be thought of as a pixel-wise classification problem determining the pixel is foreground or not. To solve this problem with deep learning, Fully Convolutional Network(FCN) [16], which removed a fully connected layer from the network, showed that astounding results at semantic segmentation. *Hyeonwoo Noh et al.* [17] established encoder-decoder architecture for image segmentation, U-Net [6] added a skip connection from encoder to decoder. FPN [18] adopted feature pyramid network inspired from multi-scale image pyramid. Mask R-CNN [19] tried segmentation with the transformed head from Faster R-CNN. As such, there are many attempts to incorporate deep learning into various fields of computer vision.

### 2.1.2 Convolutional Layer

The convolutional layer is a filter consisting of learnable parameters. It has a small receptive field, but it extends through the channels of input. The filter-sized patch is convolved across the input area. Then element-wise multiplication is processed between the parameters inside a patch and the overlapping area of an input. In this way, the CNN learns filters to activate when it detects

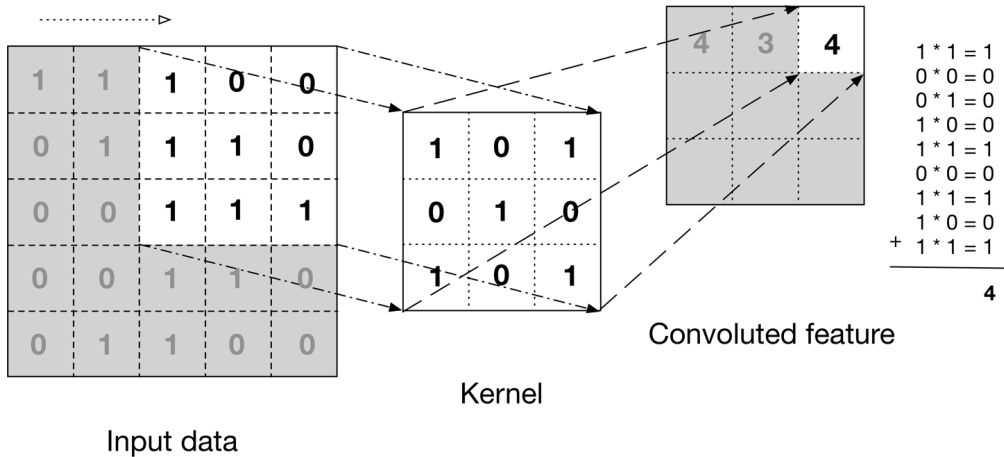


Figure 2.1: Operation of convolutional layer [20]

features at several spatial locations in the input. We can denote the convolution operation as  $*$ . For example, given the 2-dimensional image  $I$  and the filter  $K$ , the discrete convolution of the image  $I$  with filter  $K$  is given by:

$$(I * K)_{r,s} = \sum_{u=-w}^w \sum_{v=-h}^h K_{u,v} I_{r-u,s-v} \quad (2.1)$$

where  $r, s$  is the discrete coordinate of 2-dimensional space of the image and  $w, h$  is the height and width of the filter. Moreover, a convolutional layer computes the output  $x^{l+1}$  with the given input  $x^l$  as:

$$x_j^{l+1} = \sum_{i \in M_j} (K_{ij}^{l+1} * x_i^l) + B_j^{l+1} \quad (2.2)$$

where  $B^l$  denotes a bias matrix and  $M$  is the size of the filter bank. In the CNN, we stack the convolutional layers, and each output of the layer could become larger and complex. It makes a deeper layer be able to represent a sophisticated feature of the input image. Finally, the condensed feature can be used to classify the object or segment part of the image.

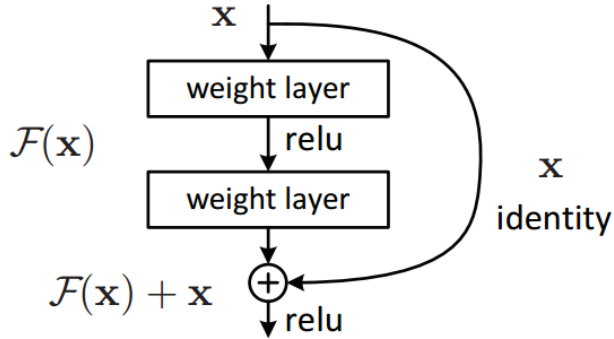


Figure 2.2: Residual block

### 2.1.3 Residual Connection

*Kaiming He et al.* [11] presented a residual connection in order to resolve vanishing/exploding gradient problem. Fig. 2.2 shows a residual block with the residual connection. As can be observed, a residual connection connects the original input  $x$  to the output of layer. Let the input is  $x^l$  and the output is  $x^{l+1}$ , the  $\mathcal{F}$  means the operation inside a block, then the residual block can be represented as:

$$x^{l+1} = \mathcal{F}(x^l) + x^l \quad (2.3)$$

Likewise, we can get the partial derivative of Eq. 2.3 for backward propagation.

$$\frac{\partial E}{\partial x^l} = \frac{\partial E}{\partial x^{l+1}} \left( 1 + \frac{\partial}{\partial x^l} \mathcal{F}(x^l) \right) \quad (2.4)$$

As shown in Eq. 2.4, the deep layer can directly backpropagate to the shallow layer. Also, the differential value of the layer must have more than one. It forces the gradient always to be one or more and addresses the vanishing/exploding gradient problem. Therefore, we can make a model with layers deep and deeper.

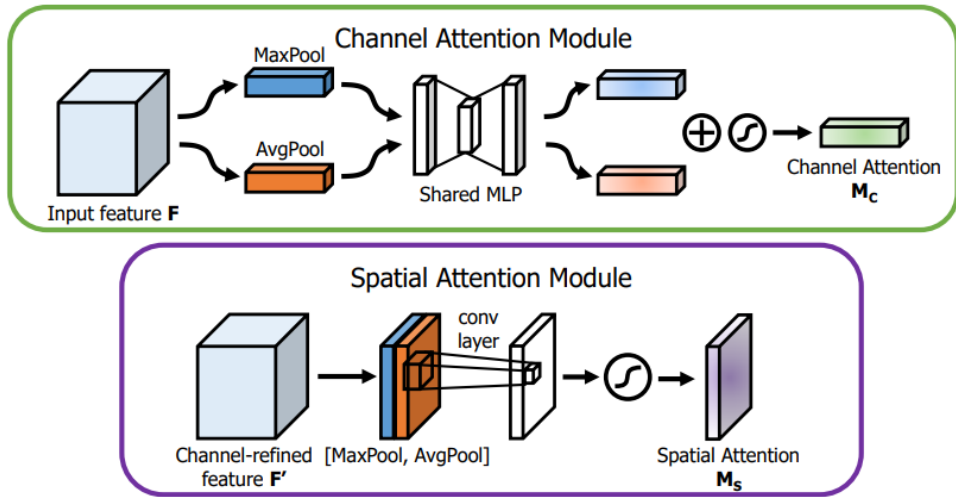


Figure 2.3: Diagram of channel attention module and spatial attention module

## 2.2 Visual Attention Mechanism

The basic idea of the attention mechanism is focusing on a specific feature. Notably, attention is a mechanism by which a network can weigh features by their importance. The attention mechanism was invented for translation in the natural language field and has been researched still actively in the deep learning field. To memorize long source sentences in translation, a vector named “*Attention Score*” used for dot product with a context vector to decide where to pay attention more. Particularly, the “*Attention Score*” vector not only indicates the importance of the element in the input vector but also is a learnable parameter during model training. In other words, the attention vector can teach the model which elements it should pay more attention to improve model performance.

The attention mechanism for CNN widely used for visual question answering or image captioning. The intention of the visual attention mechanism is simply to imitate the visual recognition system of humans. For example, when human

eyes are focusing on the object, the backgrounds are blurred. Similarly, the visual attention enhances the output of the receptive field around the objects. *Xu et al.* [21] showed through visualization how the model pays attention automatically to objects in the image. *Wang et al.* [22] stacked the attention modules for image classification. SE block[23], a kind of attention vector, can recalibrate feature maps. Based on [22][23], Bottleneck Attention Module(BAM) [24] and Convolutional Block Attention Module(CBAM) [25] was proposed. BAM and CBAM were able to easily combine with the prior CNNs and achieved higher performance. Fig. 2.3 shows channel attention and spatial attention modules from CBAM. The channel attention module weighs the importance of features at the channel level, and the spatial attention module encodes the spatial location of objects in feature maps. However, BAM and CBAM are vulnerable to get shape-prior features since max and average pooling enfeebles the detail of shape feature to reinforce spatial location information.

# Chapter 3

## Related Works

### 3.1 CNNs for Biomedical Image Segmentation

While CNNs have made groundbreaking performance in the computer vision field, to incorporate deep learning into medical imaging many studies have been conducted. The problem is that it is difficult to generalize CNN models to the whole medical dataset because gathering and annotating medical images is a fatiguing and expert task. Even more, Medical images from digital systems, usually CT, MRI, and US, are 3-dimensional data that means high computing power is needed to process large medical datasets. Nevertheless, *Suk and Shen* [26], *Suk et al.* [27], and *Pils et al.* [28] adopted CNNs to medical imaging analysis and classified patients as having Alzheimer's disease based on brain MRI.

Moreover, deep learning techniques have improved the quality of medical image segmentation. As mentioned in chapter 2.1, image segmentation is the voxel-wise classification problem in 3-dimensional medical image.(Fig. 3.1)

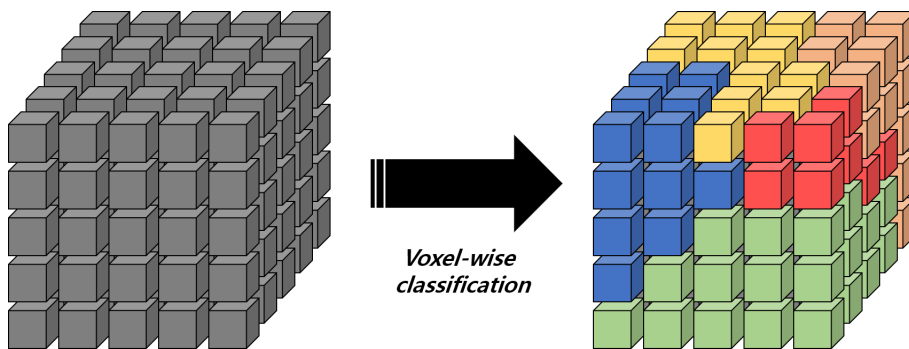


Figure 3.1: 3-dimensional image segmentation

**U-Net** U-Net [6] is the symbolic encoder-decoder CNN architecture. The U-shape model is shown in Fig. 3.2, which consists of downsampling layers as the encoder and upsampling layers as the decoder. In the encoder section, a model try to learn the coarse feature maps to capture contextual information. Subsequently, in the decoder section, the coarse feature maps are merged with skip connections to generate a fine-level dense prediction. Because of the good performance and efficient usage of GPU memory, U-Net is commonly used in various fields of biomedical image segmentation. *Çiçek et al.* [29] expanded an idea of U-Net architecture to 3-dimensional volume data.

**V-Net** Milletari et al. [30] also proposed 3D U-Net architecture named V-Net with Dice similarity coefficient(DSC) score based loss function. The DSC based loss function(Dice loss) can address the problem of data imbalance between foreground and background voxels so that later networks could use it to increase model performance.

**AutoCENet** *Chung et al.* [31] proposed CNN named *AutoCENet* to segment the liver from abdomen CT images. Since the liver is closely adjacent to other organs, they used shape-prior information and edge features to train the auto-context model. V-transition is introduced to enhance the performance which has

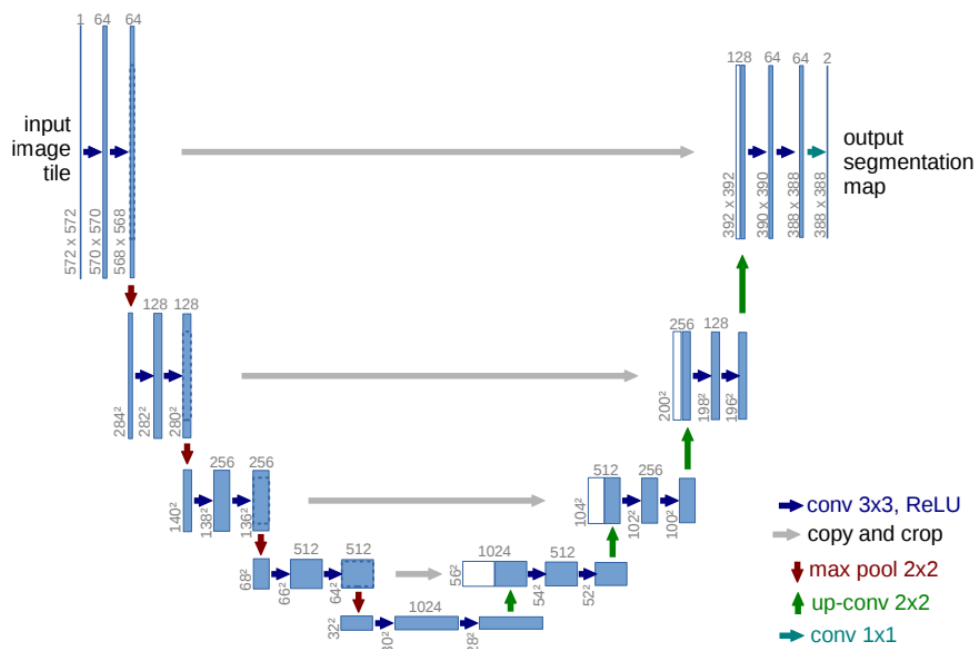


Figure 3.2: U-Net architecture



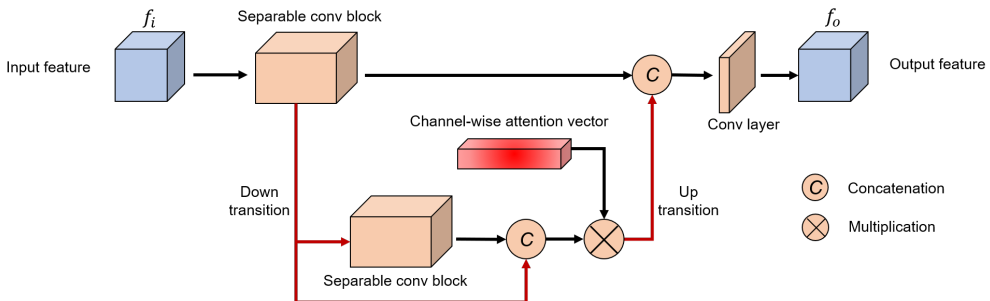


Figure 3.3: V-transition layer from AutoCENet. [31] A separable convolutional block is composed of separable convolutions, batch normalization, and element-wise non-linearity function.

few parameters but shows accurate feature maps using separable convolutions and up/down transitions.

In addition, *Gibson et al.* [32] and *Oktaay et al.* [33] tried to help CNNs learn shape-prior information. DenseVNet [32] used the additional learnable parameters to capture explicit spatial prior. The shape-prior information may be suitable for single organs with regular shape, however, not much helpful for multi-organ with various shapes such as the heart.

## 3.2 Distance Transformation for CNN Image Segmentation

### 3.2.1 Distance Transformation(DT)

Distance transformation is a kind of representation of a binary image. [34] In a binary image, the transformation of the interested area looks similar to the original binary image, but each pixel represents a distance from a predetermined set of pixels. Typically, we use the boundary of the foreground area as a set of pixels, and in distance transform, pixels inside the foreground area have a distance to the nearest boundary. Fig. 3.4 explains the preceding.

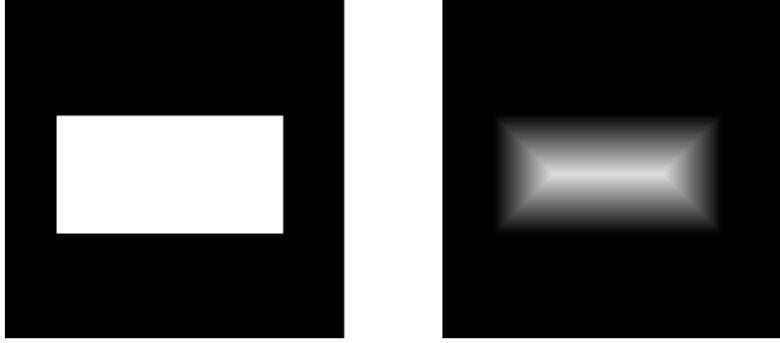


Figure 3.4: Comparison between the binary image and its distance transform. The left one shows the original rectangle binary image and the right one represents the distance transform of the left image.

A DT can be denoted as follows:

$$DT(P)[a] = \min_{y \in P} d(a, b) \quad (3.1)$$

where  $P$  is a predetermined point set such as boundary of foreground, or background, and  $a, b$  are points.  $d(a, b)$  means the function that calculates the distance between  $a$  and  $b$ . For instance, in 2-dimensional space,  $d(a, b)$  can be one of the following items:

- **Euclidean distance**  $d(a, b) = \sqrt{(a_x - b_x)^2 + (a_y - b_y)^2}$
- **Manhattan distance**  $d(a, b) = |a_x - b_x| + |a_y - b_y|$
- **Chebyshev distance**  $d(a, b) = \max(|a_x - b_x|, |a_y - b_y|)$

### 3.2.2 DT-based Image Segmentation

Distance transformation is a highly utilized task in computer vision. A DT can be used for morphological image processing such as erosion or dilation. Moreover, exact Euclidean Distance Transformation(EDT) can help to produce

$\sqrt{2}$	1	$\sqrt{2}$
1	0	1
$\sqrt{2}$	1	$\sqrt{2}$

2	1	2
1	0	1
2	1	2

1	1	1
1	0	1
1	1	1

Figure 3.5: Distance metrics for the DT in 2-dimensional 3x3 space. The left shows Euclidean distance metric, and the center shows Manhattan distance metric, and the right shows Chebyshev distance metric.

an accurate, reversible skeleton. We can use a DT to match the pattern or non-rigid image registration. One example of the application, Chamfer Matching [35] is a classic method for comparing the similarity between two images. Not only this, a DT can help blurring images and motion planning for the robot, even pathfinding.

Since DT can represent shape information, DT has been used for segmentation in the classical method. Intuitively, we can acquire DT from the binary contour image by thresholding and morphological dilation. Subsequently, a threshold the DT again to get markers that indicate the confident area of the object. Finally, the watershed algorithm can apply to the input image with markers, then the objects of interest will be segmented. This classical method is a simple solution but highly dependent on thresholding results. It means the algorithm may make failure on sophisticated images or noisy images. However, in recent years, DT can help improve CNNs performance. *Audebert et al.* [36] used signed distance transformation(SDT) as regression target to detect clear boundary and shapes. *Wang et al.* [37] combined the deep learning and watershed algorithm to detect cells. *Navarro et al.* [38] learned the model with labels, DT from labels, and contour of label image. For cardiac segmentation, *Dangi et al.* [39] added complementary decoder to regress DT, then the DT regular-

ized the network. The limitation of this paper includes that they used auxiliary tasks to regress DT, but it weakly affected to refine the shape-aware feature. To enforce the network to learn shape-aware features, we propose the attention mechanism with DT and contour probabilistic map in the next chapter.

# Chapter 4

## Methodology

### 4.1 Overview

In this chapter, the proposed architecture for cardiac segmentation is introduced. As mentioned in chapter 3, we present a CNN model for the whole heart segmentation based on shape-aware attention, which can be trained for high generalization performance. Therefore, we present the convolutional neural model with shape-aware attention to segment the whole heart from cardiac CT images. Our novel CNN used the contour image and DT, which guide the model to suppress false positive responses and focus on object boundary details.

### 4.2 CDA-Net Architecture

In order to segment the whole heart from cardiac CT images using deep learning technique, we designed a novel CNN named *Contour & Distance transform guided Attention NETWORK*(CDA-Net). The proposed architecture is shown in Fig. 4.1. Our new-fashioned CNN is based on U-Net [6] [29] structure. Based on

the backbone network, the contour transition network(CTN) and the distance transform transition network(DTTN) are added into CDA-Net. The underlying design principles for CTN and DTTN are to predict the contour probability of each voxel and coarse shape-prior information, respectively. To solve the multi-class problem, class-wise contour and DT were used as target objective functions for CDN and DTTN. The role of CTN and DTTN is refining the shape-aware features of cardiac structures. From the encoder part of the backbone, low-level features are fed to CTN to generate contour probabilistic map features and high-level features are connected to DTTN to regress DT features. Because low-level features are sufficient to generate a contour map, and DT needs high-level features that indicate the shape information of objects. As shown in Fig. 4.1, each transition block represents V-transition. [31] Afterwards, we integrated these features and the outputs of the backbone to apply attention mechanism. In the shape-aware attention module, the DT features suppress the feature responses in the background area, allowing the model to predict without false-positive responses. Also, the contour features serve to force the model to focus on the detail of the object boundary. Moreover, the penalty energy makes CTN and DTTN generate a noiseless prediction. Finally, another V-transition is applied to aggregate and refine features for multi-class segmentation results.

The details for the proposed components are described in the following subsections.

#### **4.2.1 DT Regression for Image Segmentation**

The DTTN used the DT of the ground truth binary image as the target. As mentioned in chapter 3.2, the DT demonstrates the shape information of objects similar to morphological skeletonization. DT is expected to help the model to learn multi-object shape-prior features while regressing the DT features.

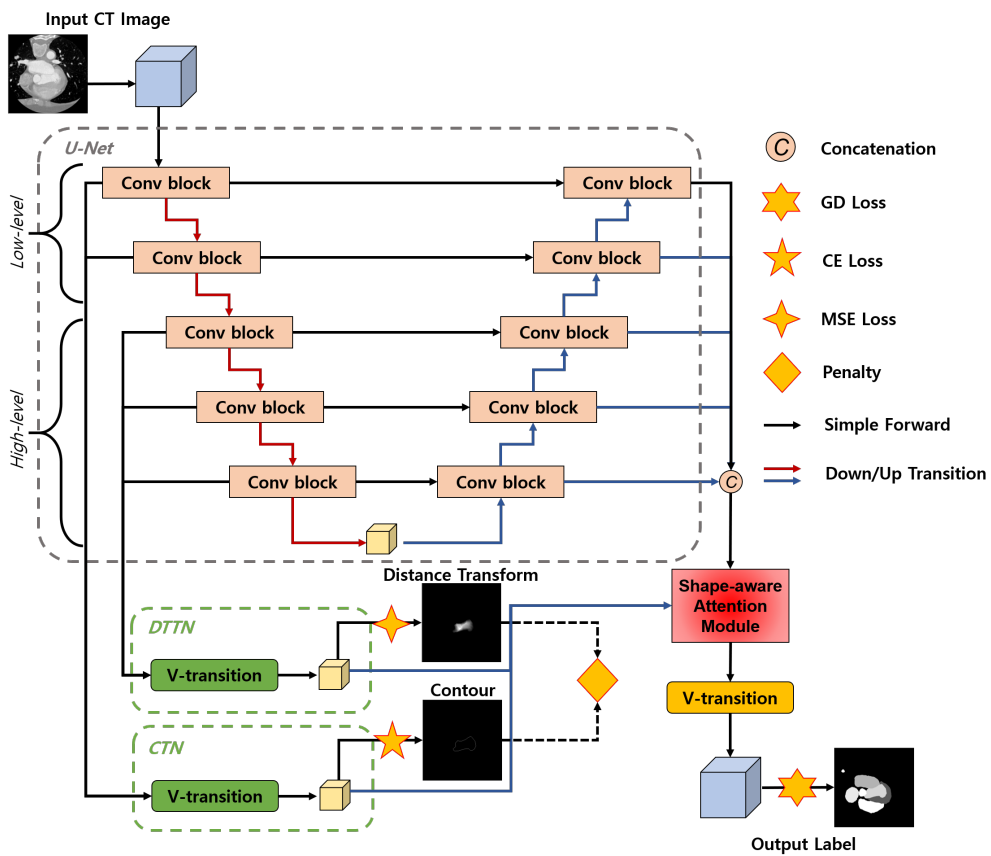


Figure 4.1: Contour & Distance transform guided Attention NETWORK

Additionally, DT can suppress the false-positive responses and generalize the model by guiding shape-prior information. In the architecture, the foreground distance transform(FDT)(shown in Fig. 4.4b) was used to represent the shape of objects.

### 4.2.2 Shape-aware Attention

The attention mechanism is used in our shape-aware attention module block. The attention block needs the input feature map with the contour probabilistic map from CTN, DT feature map from DTTN. We named the attention block as ‘‘Contour & Distance Transform Guided Shape-aware Attention’’. Fig. 4.2 shows the detail of our shape-aware attention block. Input feature is concatenated with contour feature and DT feature. Afterward, we generate the attention map by employing series of convolutional layers and a sigmoid function. The attention map is lastly multiplied element-wise(operator  $\otimes$ ) with the input feature. The output feature response is now stronger in the object area than in the background area of the feature map. We can represent  $f_o$  the output of our shape-aware attention as:

$$f_o = f_i \otimes A \tag{4.1}$$

where  $A$  is the attention map, and it is defined as:

$$A = \sigma(\mathcal{C}(f_i; f_e; f_{dt})) \tag{4.2}$$

where  $\sigma(\cdot)$  denotes the non-linear sigmoid activation function,  $\mathcal{C}(\cdot)$  is the convolutional layer, and  $(A; B)$  means the concatenation of  $A$  and  $B$ . Our attention makes the model predict the exact probabilistic map from the coarse and imprecise estimation of the backbone. Because this attention mechanism can reduce false-negative responses on feature maps and refine the feature maps with object boundary area.



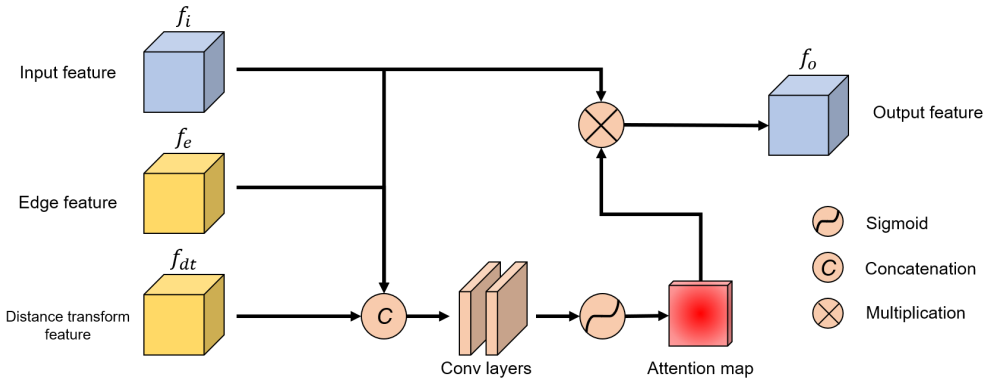


Figure 4.2: The contour & distance transform guided shape-aware attention module

### 4.2.3 Regularization with the Penalty Term to Refine Feature

The output of CTN and DTTN may produce a coarse contour probability map and DT features. Since more precise features can build a fine attention map at a shape-aware attention module, we suggest the penalty term regularize the output of CTN and DTTN. To regularize feature map responses, first, apply the sigmoid function to contour features, then each voxel of features represents the contour probability in  $[0, 1]$  values. Clamp DT features to distinguish the foreground and background voxels. Afterward, reverse the DT features by subtracting from 1 to flip foreground and background. Finally, the penalty energy for regularization is the multiplication of contour features and subtracted reverse DT features. The penalty energy takes the form:

$$E_p = \sigma(f_e) \otimes (1 - clamp(f_{dt}, 0, 1)) \quad (4.3)$$

where  $clamp(f, a, b)$  function keeps only the values in  $f$  within the range  $[a, b]$ . This penalty term is 0 if the responses of contour features don't appear in the background area of the DT features. Fig. 4.3 shows the cases of the penalty en-

ergy. Thus, it can reduce the false-positive responses from contour probabilistic map, and help the model finding the exact boundary of objects.

### 4.3 Overall Loss Function

For cardiac segmentation, as mentioned above, contour map and DT are used to formulate the loss function to predict precise results from model. The detail of our loss function is described in following. First, data imbalance problem between foreground and background voxel should be addressed in segmentation. Therefore, we used Dice loss [30] to formulate our loss function. To be exact, to address multi-class segmentation problem, generalized Dice similarity coefficient score loss(GDL) [40] was applied. Given the reference image  $R$  with voxel values  $r_n$  and the predicted probabilistic map  $P$  with elements  $p_n$ , GDL takes the form:

$$L_{GDL} = 1 - 2 \frac{\sum_{c=1}^{N_{cls}} w_c \sum_n r_{cn} p_{cn}}{\sum_{c=1}^{N_{cls}} w_c \sum_n r_{cn} + p_{cn}} \quad (4.4)$$

where  $N_{cls}$  is the number of classes to segment, and  $w_c$  is used to give weight for each label.  $w_c$  is usually determined by the number of class voxels, denotes:  $w_c = 1/(\sum_{n=1}^N r_{cn})^2$ .

To guide edge feature map for CTN, we added the loss term between the contour probabilistic map from the model and the ground truth contour images. Because the model predict the contour probability per voxel for each class, we applied Cross-Entropy loss(CEL) to each voxel.

$$L_C = - \sum_{c=1}^M y_c \log(p_c) \quad (4.5)$$

where  $M$  is the number of classes,  $y$  is the binary indicator if class label  $c$  is the correct classification, and the  $p$  denotes the class probability. Particularly, only 2 classes used for contour, 0 means the background and 1 indicates the contour,

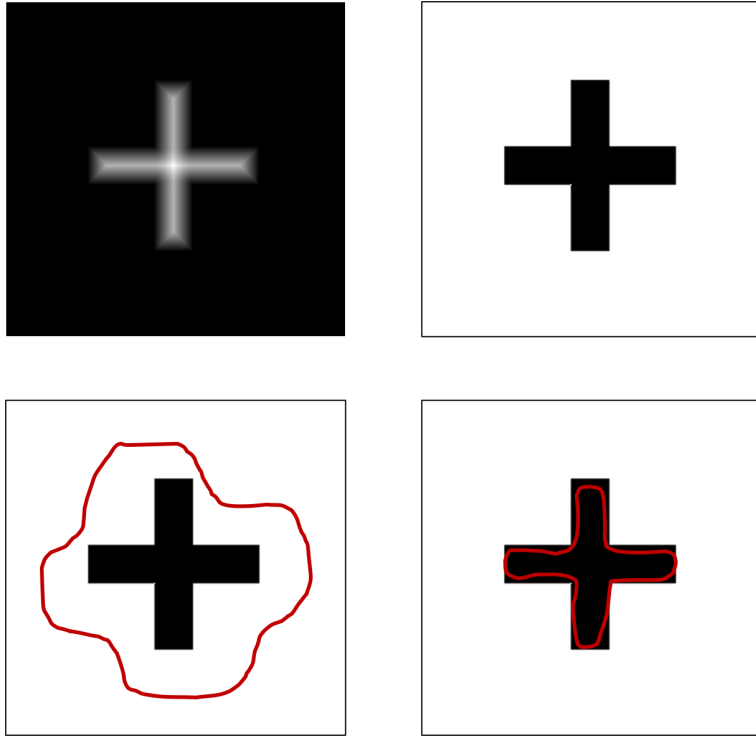


Figure 4.3: The example case of penalty energy to refine contour and DT feature. White means a higher value, and the red line means the predicted contour. The upper left is the DT feature map. The upper right image represents clamped reverse image of the DT feature. The bottom images show overlapping the contour map to the DT feature. The bottom left image will have a heavy penalty since the contour features are overlapping background area which has high values. Likewise, the bottom right image will have small penalty energy. The penalty energy plays a role that the contour feature responses do not appear in the background to find exact boundary line.

Cross-Entropy Loss can be replaced by Binary Cross-Entropy loss(BCEL).

$$L_C = -(y\log(p) + (l - y)\log(l - p)) \quad (4.6)$$

In addition, since there exists only few contour voxels compared to the background voxels, we have set the weights [0.001, 0.999] to solve highly imbalanced data problem. Then, the contour loss function can directly affect to regress contour features.

Likewise to refine DT feature for DTTN, Mean Square Error loss (MSE) is combined to our loss function. The MSE loss for predicted DT map  $P$  and the ground truth  $Y$  with the number of voxel  $n$  is

$$L_{DT} = \frac{1}{n} \sum_{i=1}^n (y_i - p_i)^2 \quad (4.7)$$

As shown from equation, MSE loss makes the big difference in voxel value even greater. In other words, it enforce the model finding the center of shape more rather than finding boundary of shape. Finally, we formulated our loss function to minimize for CDA-Net. We added all above term with weights. Subsequently, we merged two term more, first,  $E_p$  the penalty term for the edge feature and DT feature, and second,  $L_{DSV}$  deep supervision loss for intermediate feature maps.  $L_O$  means the loss for final probabilistic map output. GDL is used for  $L_O$  and  $L_{DSV}$ . Our final loss function is following:

$$L = \lambda_1 L_O + \lambda_2 L_{DSV} + \lambda_3 L_C + \lambda_4 L_{DT} + \lambda_5 E_p \quad (4.8)$$

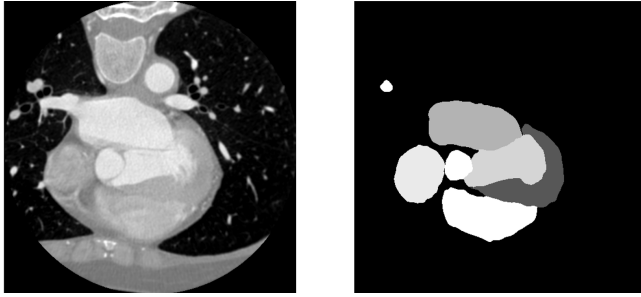
In equation,  $\lambda$  means the weight of each term. We used  $\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 20, \lambda_4 = 10, \lambda_5 = 1$  in the experiments.

#### 4.4 Data Preparation

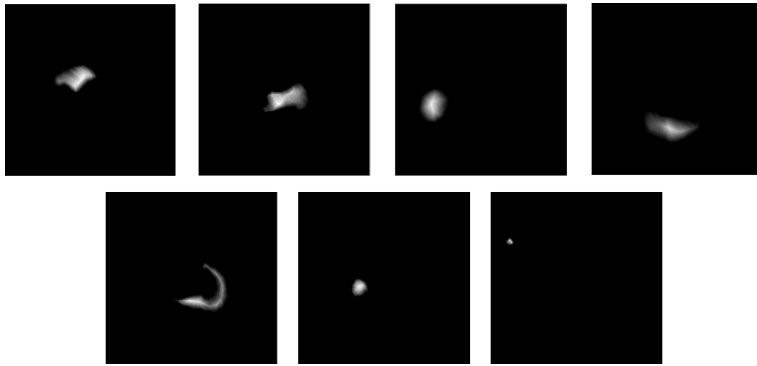
For CTN and DTTN, the contour images and DTs of ground truth labels are necessary for training. Fig. 4.4 shows the example input images to train the

model. The contour images(Fig. 4.4c) were also used to give detail of boundary feature. To acquire the ground truth contour image, the Prewitt filter is applied for a 3-dimensional direction to generate a gradient image. The FDT(Figure 4.4b) was compute by the linear time algorithm [41] for describing the shape-prior of object. In the ground truth FDT, the higher value indicates the center of the object. Both the contour image and FDT of ground truth are produced class by class to segment multi-class.

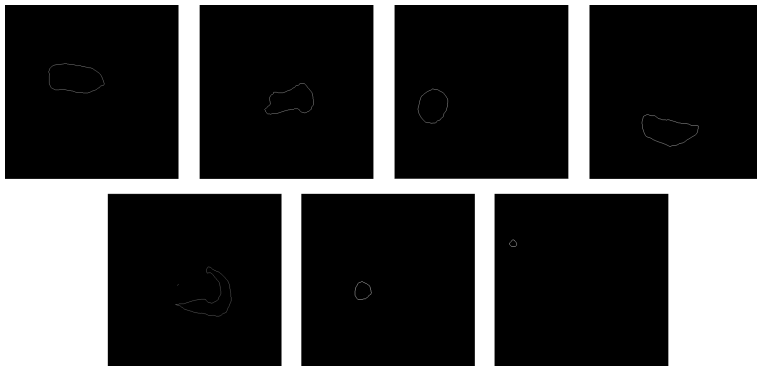
Because of memory limitation, we resized the image to 128x128x64 size for the model input. The input images were pre-processed with fixed windowing values between -300 to 1000. We also normalized the image voxel values into [0,1]. To generalize the model, Gaussian noise, rotation, and cutout were used randomly for data augmentation.



(a) Input CT image and ground truth label image



(b) DTs of ground truth label image



(c) Contour images of ground truth label image

Figure 4.4: The example of input images. Each label has the DT and contour image.

# Chapter 5

## Experimental Results

### 5.1 Dataset

We learned our model with Multi-Modality Whole Heart Segmentation 2017(MM-WHS 2017) dataset [42] [43], which provides 60 CT and 60 MRI cardiac images. Each set of 60 images is included with 20 images with label-annotated data for training and 40 images for testing. We used only the CT dataset for validation. The WHS ground truth data were manually labeled by well-trained students majoring in biomedical engineering or medical physics.[42]. Seven substructures were selected as interest area in the WHS study, including:

- 1) LV: the left ventricular cavity
- 2) RV: the right ventricular cavity
- 3) LA: the left atrial cavity
- 4) RA: the right atrial cavity
- 5) Myo: the myocardium of left ventricle
- 6) AA: the ascending aorta trunk

7) PA: the pulmonary artery trunk

The training dataset was split into a training/validation set to generalize the model using n-fold cross-validation.

## 5.2 Evaluation Metrics

Evaluation of the segmentation results was conducted with the Dice similarity coefficient score and Jaccard index. Given the binary labeled masks  $X$  and  $Y$ , Dice similarity coefficient(DSC) and Jaccard index(JI) as follows:

$$DSC(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|} \quad (5.1)$$

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|} \quad (5.2)$$

We also evaluate the surface distance metrics, 95% Hausdorff distance(HD) and average symmetric surface distance(ASSD), to demonstrate the performance of the proposed shape-aware attention module. The 95% HD processed without 5% of outlying voxels since it is more robust by avoiding noisy outliers. HD is defined as follows:

$$HD(X, Y) = \max\left\{\max_{s_X \in S_X} d(s_X, S_Y) + \max_{s_Y \in S_Y} d(s_Y, S_X)\right\} \quad (5.3)$$

where  $d(p, S_X)$  means the shortest distance from an arbitrary voxel  $p$  to a set of surface  $S_X$ .

$$d(p, S_X) = \min_{s_X \in S_X} \|p - s_X\|_2 \quad (5.4)$$

We can also define the distance function between two sets:

$$D(S_X, S_Y) = \sum_{s_X \in S_X} d(s_X, S_Y) \quad (5.5)$$



Then the ASSD can be defined as follows:

$$ASSD(X, Y) = \frac{1}{|S_X| + |S_Y|} (D(S_X, S_Y) + D(S_Y, S_X)) \quad (5.6)$$

To prove the reduction of false-positive responses, sensitivity and precision are used:

$$S = \frac{TP}{TP + FN} \quad (5.7)$$

$$P = \frac{TP}{TP + FP} \quad (5.8)$$

where TP, FP, and FN are the numbers of true positive, false positive, and false negative output voxels.

### 5.3 Experiments Detail

We attempted to train our model to minimize loss  $L$ (Eq. 4.8). While training the network, we used the Adam optimizer with a learning rate of 1e-3. The segmentation outputs are obtained by applying the softmax function to the final feature maps. All training was conducted on Intel 10 core 19-7900X processor and 24GB Nvidia Titan RTX GPU machine with 128GB memory. We implemented the proposed network using the PyTorch framework.

### 5.4 Comparison with State-of-the-Art

**Quantitative Results:** Table 5.1 and Table 5.2 show the quantitative results of cardiac segmentation. Dice similarity coefficient score and Jaccard Index were computed for all cardiac substructures including LV, RV, LA, RA, LV-myocardium, AA, PA, and whole heart(WH). We evaluated the performance of our model by comparing the other state-of-the-art models. U-Net[29], VoxResNet[44], DenseVNet[32],

and Attention Gated U-Net(AGU-Net)[33] are compared to our proposed network. As can be seen, CDA-Net outperforms the other state-of-the-art networks in terms of DSC and JI. Table 5.1 and Table 5.2 shows that the attention mechanism can enhance the performance finding uneven boundary structures i.e. LV-myocardium, AA, PA. In addition, the overall results present the superiority of the proposed network.

**Qualitative Results:** Fig. 5.1 shows the axial slices of the cardiac segmentation results from models. The results show that our model outperforms the other networks. Notably, our model predicted the exact boundary of cardiac substructures without any false-positive responses. Fig. 5.2 displays the volume and surface of predicted labels. Fig. 5.2 indicates that our model predicts less noisy and smooth surface results. In contrast, the other results exist mis-segmented outputs and imprecise boundary of substructures. DenseVNet and AGU-Net reduces false positive and predicts less rough surface than VoxResNet, UNet, but still cannot find the boundary between LV(Red) and RV(Skyblue).

## 5.5 Ablation Study

In this subsection, we prove the effectiveness of our proposed shape-aware attention module and the penalty term with following ablation studies. we validate the performance by conducting following experiments. Table 5.3, 5.4, 5.5, 5.6, and 5.7 are showing the contribution of each component. The baseline in the ablation study is the composition of backbone U-Net [6] and a single V-transition [31]. From the baseline, five additional experiments were studied: with self-attention(base+CBAM), with only CTN(base+CTN), with only DTTN(base+DTTN), without the penalty term(base+CTN+DTTN), and the proposed network(base+CTN+DTTN+penalty). The base+CBAM network replaced the shape-aware attention module with CBAM.

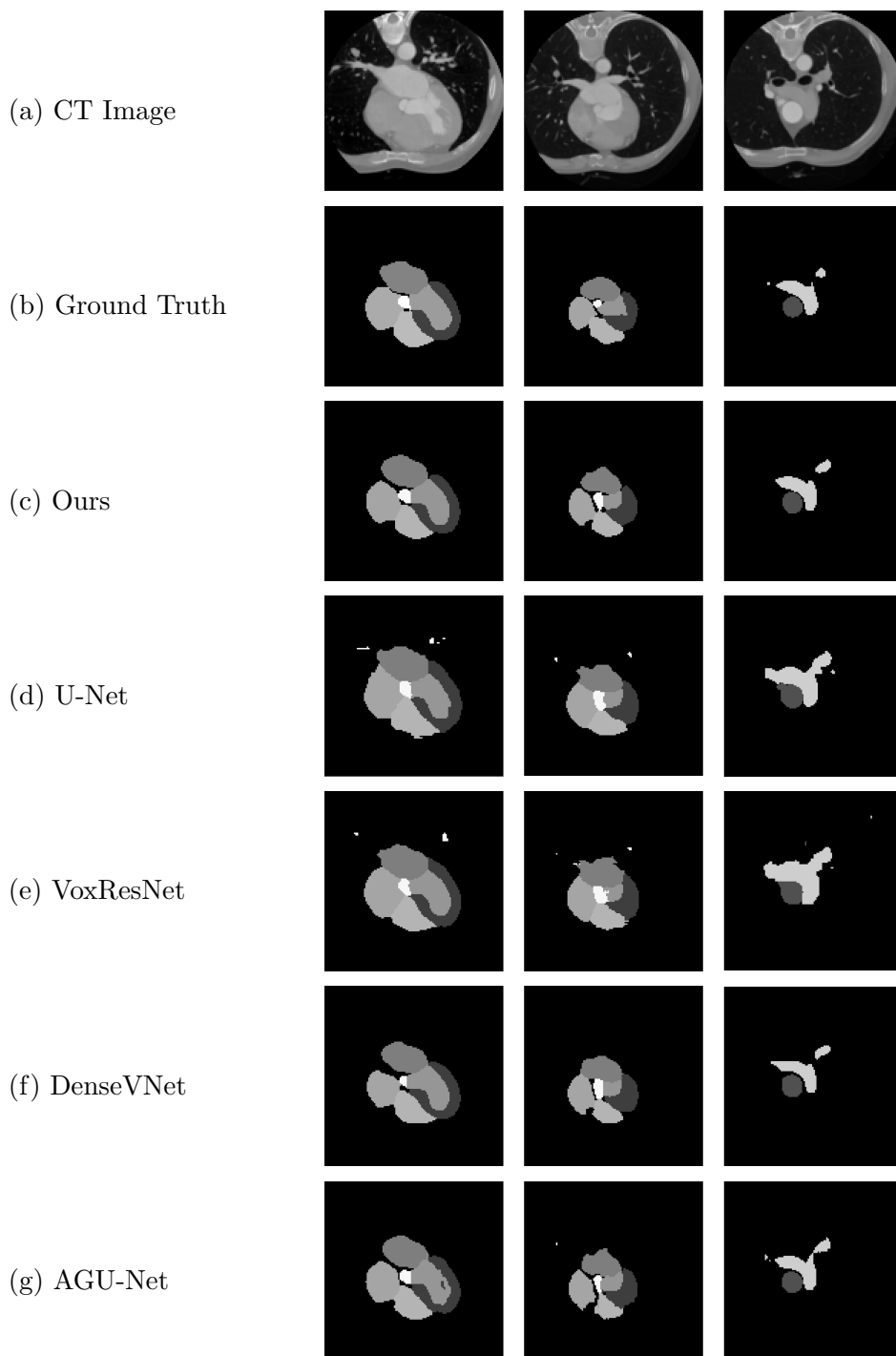


Figure 5.1: The axial slices of segmentation results

Network	LA	LV	RA	RV	LV-myoe	AA	PA	WH
U-Net	0.8951	0.8347	0.8172	0.7712	0.7845	0.8081	0.7111	0.8184
VoxResNet	0.8750	0.7880	0.7197	0.7328	0.7461	0.7525	0.6062	0.7655
DenseVNet	0.8708	0.7998	0.8288	0.7602	0.7921	0.7716	0.6810	0.8100
AGU-Net	0.8575	0.7921	0.8448	0.7950	0.8017	0.8142	0.7180	0.8203
Ours	<b>0.9005</b>	<b>0.8656</b>	<b>0.8935</b>	<b>0.8638</b>	<b>0.8242</b>	<b>0.8736</b>	<b>0.7915</b>	<b>0.8700</b>

Table 5.1: Dice similarity coefficient score of CDA-Net and the state-of-the-arts

Network	LA	LV	RA	RV	LV-myoe	AA	PA	WH
U-Net	0.8138	0.7170	0.6984	0.6353	0.6508	0.7049	0.5681	0.6955
VoxResNet	0.7865	0.6559	0.5735	0.5850	0.6009	0.6400	0.4503	0.6246
DenseVNet	0.7762	0.6764	0.7265	0.6315	0.6629	0.6746	0.5453	0.6866
AGU-Net	0.7693	0.6809	0.7526	0.6775	0.6806	0.7208	0.5979	0.7059
Ours	<b>0.8221</b>	<b>0.7709</b>	<b>0.8113</b>	<b>0.7657</b>	<b>0.7118</b>	<b>0.7999</b>	<b>0.6775</b>	<b>0.7733</b>

Table 5.2: Jaccard Index of CDA-Net and the state-of-the-arts

*1) Effectiveness of CTN and DTTN* The distance transform transition network(DTTN) is proposed to detect shape-prior features, and the contour transition network(CTN) is proposed to learn precise contour features. As can be observed in tables, CTN alone or DTTN alone cannot predict precise segmentation results. However, they are related complementary since CTN can detect accurate surfaces and DTTN can learn shape-prior features of objects.

*2) Effectiveness of our shape-aware attention* Compared to the self-attention module(base+CBAM), our shape-aware attention reduces false-positive responses effectively. Moreover, with contour features and DT features, the model can easily learn the shape of objects and boundary area.

*3) Effectiveness of proposed penalty term* As can be seen in 5.3, the penalty term between the contour features and DT features can refine the contour probabilistic map of CTN. In this way, the penalty term helps to predict the accurate surface of objects by generating exact contour probabilities. The

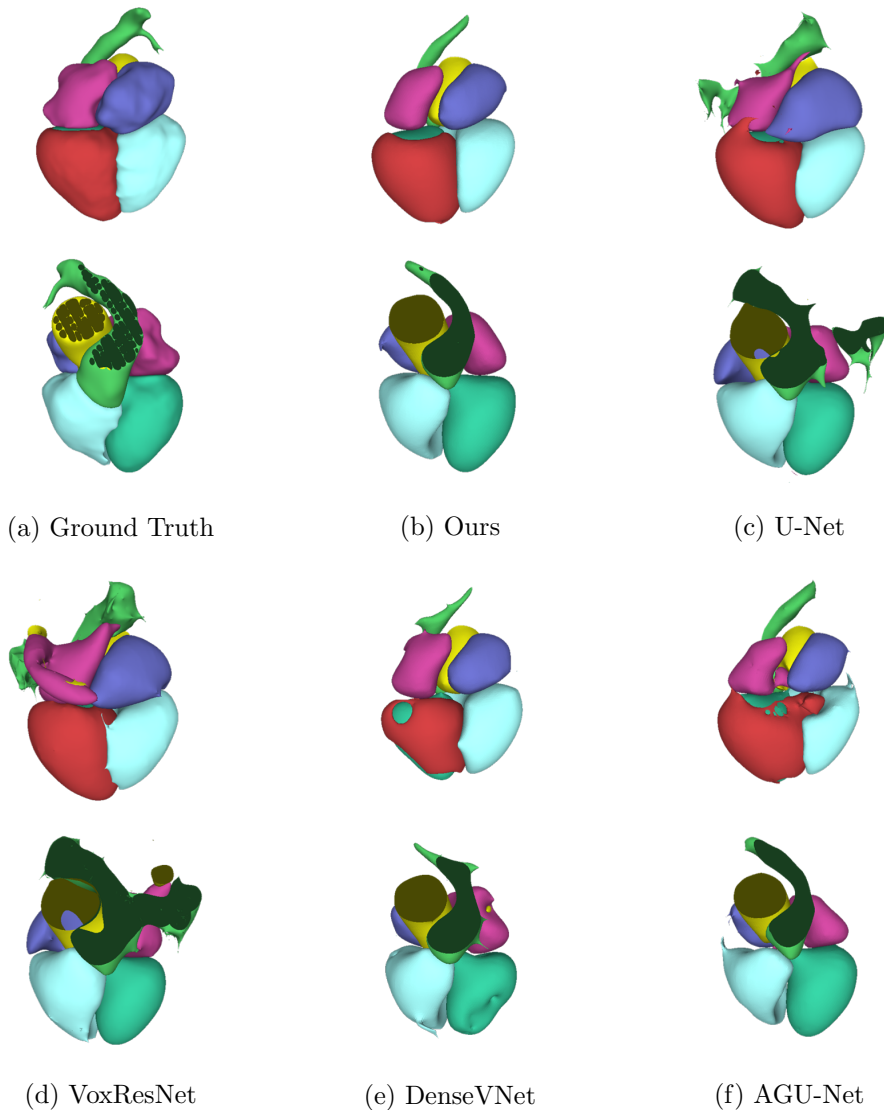


Figure 5.2: The model output visualization of cardiac structures. The second row of each image is shown without LV-myocardium.

95% HD from Table 5.4 and ASSD from Table 5.5 shows that the penalty term enforces to detect precise surface. Additionally, from Table 5.7, the model with

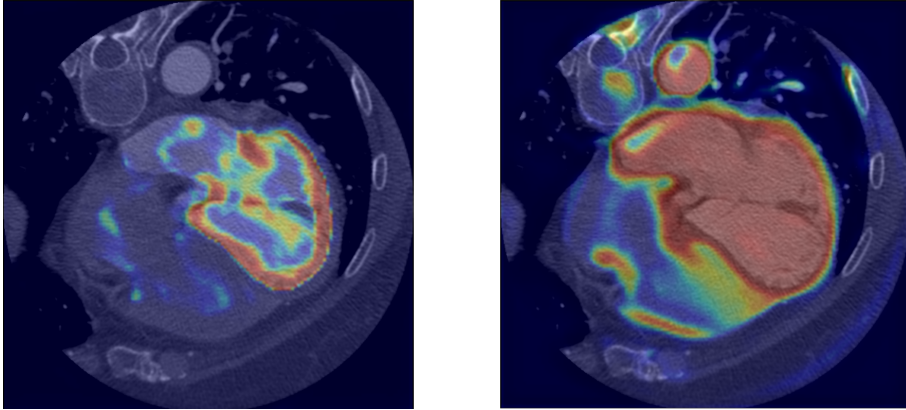


Figure 5.3: Visualization of left ventricle the contour probabilistic map of CTN. The left image is applied the proposed penalty term, and the right image is not.

the penalty term scores high precision than the model without the penalty term. In other words, the results describe that the penalty term effectively restraint the false-positive responses.

Network	LA	LV	RA	RV	LV-myo	AA	PA
base	0.861	0.906	0.819	<b>0.854</b>	0.827	0.923	0.769
base+CBAM [25]	0.838	0.896	0.740	0.821	0.817	0.917	0.793
base+CTN	0.880	0.901	0.823	0.845	0.817	0.942	0.815
base+DTTN	0.851	0.905	0.763	0.841	0.833	0.923	0.793
base+CTN+DTTN	0.893	<b>0.924</b>	0.843	<b>0.857</b>	0.842	0.946	0.800
base+CTN+DTTN+penalty(Ours)	<b>0.901</b>	<b>0.929</b>	<b>0.853</b>	<b>0.852</b>	<b>0.861</b>	<b>0.963</b>	<b>0.843</b>

Table 5.3: Dice similarity coefficient score of CDA-Net and its ablations

Network	LA	LV	RA	RV	LV-myo	AA	PA
base	6.741	2.136	8.136	5.260	2.245	3.735	11.981
base+CBAM [25]	5.190	2.379	5.899	5.407	2.436	1.796	9.212
base+CTN	3.699	2.762	6.334	4.985	2.528	1.667	8.279
base+DTTN	4.311	4.566	8.518	7.381	3.771	5.849	9.322
base+CTN+DTTN	3.560	1.814	<b>5.685</b>	4.655	2.513	1.528	8.376
base+CTN+DTTN+penalty(Ours)	<b>3.105</b>	<b>1.678</b>	6.730	<b>3.682</b>	<b>1.913</b>	<b>1.173</b>	<b>7.243</b>

Table 5.4: 95% HD of CDA-Net and its ablations

Network	LA	LV	RA	RV	LV-myo	AA	PA
base	1.432	0.942	1.804	1.453	0.801	0.798	2.411
base+CBAM [25]	1.371	0.842	1.932	1.496	0.853	0.567	1.672
base+CTN	0.997	0.887	1.642	1.371	0.828	0.472	<b>1.445</b>
base+DTTN	1.158	1.248	2.143	1.669	1.038	0.902	1.669
base+CTN+DTTN	0.962	0.647	<b>1.365</b>	1.241	0.737	0.545	1.553
base+CTN+DTTN+penalty(Ours)	<b>0.844</b>	<b>0.637</b>	1.468	<b>1.122</b>	<b>0.658</b>	<b>0.362</b>	<b>1.437</b>

Table 5.5: ASSD of CDA-Net and its ablations

Network	LA	LV	RA	RV	LV-myo	AA	PA
base	0.826	0.917	0.756	0.873	0.836	0.897	0.744
base+CBAM [25]	0.912	<b>0.955</b>	0.869	0.875	0.849	0.951	0.860
base+CTN	0.887	0.921	0.863	0.861	0.847	0.932	<b>0.865</b>
base+DTTN	0.907	0.889	0.841	0.841	0.865	0.885	0.841
base+CTN+DTTN	0.926	0.940	<b>0.876</b>	0.887	<b>0.890</b>	0.948	0.830
base+CTN+DTTN+penalty(Ours)	<b>0.932</b>	<b>0.951</b>	0.852	<b>0.918</b>	0.828	0.962	<b>0.855</b>

Table 5.6: Sensitivity of CDA-Net and its ablations

Network	LA	LV	RA	RV	LV-myo	AA	PA
base	<b>0.913</b>	0.900	<b>0.913</b>	0.847	0.825	0.922	0.819
base+CBAM [24]	0.787	0.849	0.693	0.789	0.792	0.897	0.748
base+CTN	0.881	0.894	0.815	0.840	0.802	0.958	0.786
base+DTTN	0.821	0.908	0.758	<b>0.849</b>	0.810	<b>0.971</b>	0.769
base+CTN+DTTN	0.867	0.913	0.827	<b>0.842</b>	0.805	0.950	0.792
base+CTN+DTTN+penalty(Ours)	0.878	<b>0.910</b>	0.891	0.835	<b>0.908</b>	<b>0.966</b>	<b>0.842</b>

Table 5.7: Precision of CDA-Net and its ablations

# Chapter 6

## Conclusion

Cardiac segmentation is a challenging problem because of various heart shapes and ambiguous multi-organ boundaries. To address the problem, we proposed a novel convolutional neural network named “*Contour & Distance transform guided shape-aware Attention NETWORK*”. Our intention of network design is to capture the boundary-aware features and the shape-aware features to improve the model performance. We presented the auxiliary transitions to refine the contour probabilistic map and DT features. In addition, we proposed the penalty term that suppresses the false responses on the contour probabilistic map. The penalty term made the model predict more precise contour features and DT features so that the model can easily pay attention to objects and boundary areas by minimizing the penalty term.

We applied the proposed attention mechanism to U-Net based backbone network. To validate our attention mechanism, we conducted the experiments by eliminating the components one by one. The experimental results show that the proposed method can restraint false-positive responses and outperform the



cutting-edge segmentation models by refining the shape-aware and boundary-aware attention map. The exact contour probabilistic map and DT feature lead to outstanding segmentation results. Note that our proposed network predicts high-quality segmentation results on rough surface objects. In the future, the proposed network can be adopted in vessel segmentation and tiny tissue segmentation which has unseen surface and various shapes.

# Bibliography

- [1] K. Mc Namara, H. Alzubaidi, and J. K. Jackson, “Cardiovascular disease as a leading cause of death: how are pharmacists getting involved?” *Integrated Pharmacy Research & Practice*, vol. 8, pp. 1–11, Feb 2019.
- [2] S. S. Virani *et al.*, “Heart disease and stroke statistics-2020 update: A report from the american heart association,” *Circulation*, vol. 141, no. 9, pp. e139–e596, 2020.
- [3] H. C. McGill, C. A. McMahan, and S. S. Gidding, “Preventing heart disease in the 21st century,” *Circulation*, vol. 117, no. 9, pp. 1216–1227, 2008.
- [4] M. J. O’Donnell *et al.*, “Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (interstroke): a case-control study,” *The Lancet*, vol. 388, no. 10046, pp. 761 – 775, 2016.
- [5] Wikipedia, “Heart — wikipedia, the free encyclopedia,” 2020, [Online; accessed 22-November-2020]. [Online]. Available: <https://simple.wikipedia.org/w/index.php?title=Heart&oldid=7043173>
- [6] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” 2015.

- [7] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012, pp. 1097–1105.
- [9] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.
- [10] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2015.
- [11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2015.
- [12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” 2014.
- [13] R. Girshick, “Fast r-cnn,” 2015.
- [14] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” 2016.
- [15] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” 2018.
- [16] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” 2015.

- [17] H. Noh, S. Hong, and B. Han, “Learning deconvolution network for semantic segmentation,” 2015.
- [18] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” 2017.
- [19] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.
- [20] J. Patterson and A. Gibson, *Deep Learning: A Practitioner’s Approach*, 1st ed. O’Reilly Media, Inc., 2017.
- [21] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 2048–2057.
- [22] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” 2017.
- [23] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-excitation networks,” 2019.
- [24] J. Park, S. Woo, J.-Y. Lee, and I. Kweon, “Bam: Bottleneck attention module,” 07 2018.
- [25] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in *Computer Vision – ECCV 2018*, V. Ferrari,

- M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 3–19.
- [26] H.-I. Suk, S.-W. Lee, and D. Shen, “Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis,” *NeuroImage*, vol. 101, pp. 569 – 582, 2014.
- [27] H.-I. Suk and D. Shen, “Deep learning-based feature representation for ad/mci classification,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 583–590.
- [28] S. M. Plis, D. R. Hjelm, R. Salakhutdinov, E. A. Allen, H. J. Bockholt, J. D. Long, H. J. Johnson, J. S. Paulsen, J. A. Turner, and V. D. Calhoun, “Deep learning for neuroimaging: a validation study,” *Frontiers in Neuroscience*, vol. 8, p. 229, 2014.
- [29] Özgün Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3d u-net: Learning dense volumetric segmentation from sparse annotation,” 2016.
- [30] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” 2016.
- [31] 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,” 2020.
- [32] 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 Transactions on Medical Imaging*, vol. 37, no. 8, pp. 1822–1834, Aug 2018.
- [33] O. Oktay, J. Schlemper, L. L. 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,” 2018.
- [34] G. Borgefors, “Distance transformations in digital images,” *Computer Vision, Graphics, and Image Processing*, vol. 34, no. 3, pp. 344 – 371, 1986.
- [35] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf, “Parametric correspondence and chamfer matching: Two new techniques for image matching,” in *IJCAI*, 1977, pp. 659–663.
- [36] N. Audebert, A. Boulch, B. L. Saux, and S. Lefèvre, “Distance transform regression for spatially-aware deep semantic segmentation,” 2019.
- [37] D. Shen, G. Wu, and H.-I. Suk, “Deep learning in medical image analysis,” *Annual Review of Biomedical Engineering*, vol. 19, no. 1, pp. 221–248, 2017.
- [38] F. Navarro, S. Shit, I. Ezhov, J. Paetzold, A. Gafita, J. C. Peeken, S. E. Combs, and B. H. Menze, “Shape-aware complementary-task learning for multi-organ segmentation,” in *Machine Learning in Medical Imaging*, H.-I. Suk, M. Liu, P. Yan, and C. Lian, Eds. Cham: Springer International Publishing, 2019, pp. 620–627.
- [39] S. Dangi, C. A. Linte, and Z. Yaniv, “A distance map regularized cnn for cardiac cine mr image segmentation,” *Medical Physics*, vol. 46, no. 12, pp. 5637–5651, Dec 2019.

- [40] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, “Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations,” *Lecture Notes in Computer Science*, p. 240–248, 2017.
- [41] C. R. Maurer, Rensheng Qi, and V. Raghavan, “A linear time algorithm for computing exact euclidean distance transforms of binary images in arbitrary dimensions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 2, pp. 265–270, 2003.
- [42] X. Zhuang and J. Shen, “Multi-scale patch and multi-modality atlases for whole heart segmentation of mri,” *Medical Image Analysis*, vol. 31, pp. 77 – 87, 2016.
- [43] X. Zhuang, “Challenges and methodologies of fully automatic whole heart segmentation: A review,” *Journal of Healthcare Engineering*, vol. 4, p. 981729, Jan 1900.
- [44] H. Chen, Q. Dou, L. Yu, and P.-A. Heng, “Voxresnet: Deep voxelwise residual networks for volumetric brain segmentation,” 2016.

## 초록

심장 영상 분할은 심장 의료 수술의 설계에 있어 중요한 과정이다. 최근에는 의료 영상에서 장기 분할을 위해 딥러닝을 이용한 CNN들이 많이 제안되었다. 다른 주요 인체의 장기들에 비해, 심장은 여러 개의 부분 장기들로 구성되어있다. 그리고 심장의 부분 장기들은 매우 인접하여 존재하기 때문에 CNN은 장기들의 경계선 부분에 집중하여 학습되어야한다. 본 논문에서는 심장 분할 모델의 성능을 높이기 위하여 distance transform과 윤곽선 영상을 활용하여 모양 인식 특징맵 그리고 경계선 인식 특징맵을 학습할 수 있는 네트워크 구조를 제안하였다. 또한 부분 장기들의 경계선 특징을 모델 학습에 도와줄 수 있는 모양 인식 그리고 경계선 인식 특징 attention 기술을 제안한다. 또한 정확한 경계선 확률맵을 찾기 위한 정규화 방법도 제안한다. 실험 결과들은 제안한 네트워크가 다른 최신의 영상 분할 네트워크들에 비해 4.97% 높은 DSC 성능을 보여줌으로 영상 분할 결과가 더욱 정확함을 알려준다. 영상 분할 네트워크의 학습과 검증에 있어 20개의 심장 CT를 사용하였고, 40개의 영상을 실험에 사용하였다. 이에 더해, 제안한 네트워크의 분할 결과는 정확한 경계면과 distance transform을 얻는 것이 모델의 성능을 높여 줄 수 있음을 보여준다. 본 논문에서는 제안한 모양 인식 attention 방법의 검증과 중요성을 강조하기 위해 ablation study를 진행하였다.

**주요어:** 인공 신경망, 모양 인식 어텐션, 의료 영상 분할, 심장 분할

**학번:** 2019-23191