SEMANTIC SEGMENTATION OF CARDIAC MRI WITH WEAK SUPERVISION AND DOMAIN KNOWLEDGE

An Honors Thesis Presented

By

JIANXIONG LIN

Approved as to style and content by:

**** Madalina Fiterau Brostean 05/25/23 14:53 ****
Chair

**** Evangelos Kalogerakis 05/25/23 14:56 ****
Committee Member

**** Philip Sebastian Thomas 05/25/23 15:40 ****
Honors Program Director

ABSTRACT

Deep learning models have proven effective for image segmentation tasks in the medical field, particularly in identifying organs or lesions within anatomical images, and often outperform other segmentation methods. However, data sparsity can hinder model training since acquiring ground-truth masks for medical images is a challenging and time-consuming task. To address this issue, weak supervision can be used to train models with weak labels and domain knowledge. In this paper, we propose a weakly supervised segmentation strategy that integrates domain knowledge to achieve comparable segmentation results to supervised models. Our method utilizes a custom loss function that incorporates size and shape constraints using domain knowledge. To enforce the shape constraint, we train a representation encoder with synthesized masks in an unsupervised manner, and penalize the model if the l2 distance of the segmentation output is greater than the radius of the representation hyper-sphere. To enforce the size constraint, penalize the model if the size of the segmentation output is outside the bound which is determined by the domain knowledge. Our model is evaluated on the Left Ventricular Segmentation Challenge (LVSC) dataset, achieving a dice score of 0.809, 0.67, 94.16 on the testing set and trained with the partial, weak, ground-truth labels, respectively. Our contributions include introducing weak constraints that enable deep learning models to leverage prior size and shape information and proposing a method for segmenting cardiac MRI data with no annotations while producing comparable results to supervised models.

1 Introduction

Heart disease is the leading cause of death in the United States. According to the CDC, in 2020, about 697,000 people in the United States died from heart disease—that's 1 in every 5 deaths. Common heart disease is coronary artery disease or heart attack, and congenital heart defect (CHD) is a life-born defect that affects a portion of our population and about 1 in 4 newborn babies with severe CHD. Irregular heartbeat increased the risk of infection in the heart muscle, and weakness in the heart leads to life-threatening situations. The treatment of heart disease depends on the severity and type of the heart disease, and the effective way to analyze and examine the heart or internal organ is through magnetic resonance imaging (MRI). In the diagnosis stage, cardiologists are using MRI scans to examine the functionality and the structure of the heart and provide patients with effective treatments.

From the MRI scan, cardiologists are able to look at the important piece of information to examine the functionality of the heart: end-diastolic volume (EDV) and end-systole volume (ESV). These two pieces of information allowed physicians to calculate the percentage the blood that leaves the heart (ejection fraction) and the volume of blood being pumped out from the left ventricle (stroke volume) which are crucial in determining the condition of the heart. A diagram of the heart is shown in figure 1.

However, the manual work to analyze and segment a patient's time series of MRI scans is too time-consuming and prone to human error. In recent years, researchers developed a new way to accomplish this task by deploying deep learning and computer vision which could save cardiologists tremendous time in manually calculating EDV and ESV. The deep learning model is the state-of-art model for performing image classification or segmentation. Convolutional Neural Network (CNN) is one of the fundamental neural networks that could be used in this task. CNN consists of input, hidden, and output layers. One or more of the hidden layers will perform the convolutions. The convolution is performing the dot product between the convolutional kernels and the input matrix. Through the convolution operation, it is able to extract the important features from the images such as edges and shapes, and

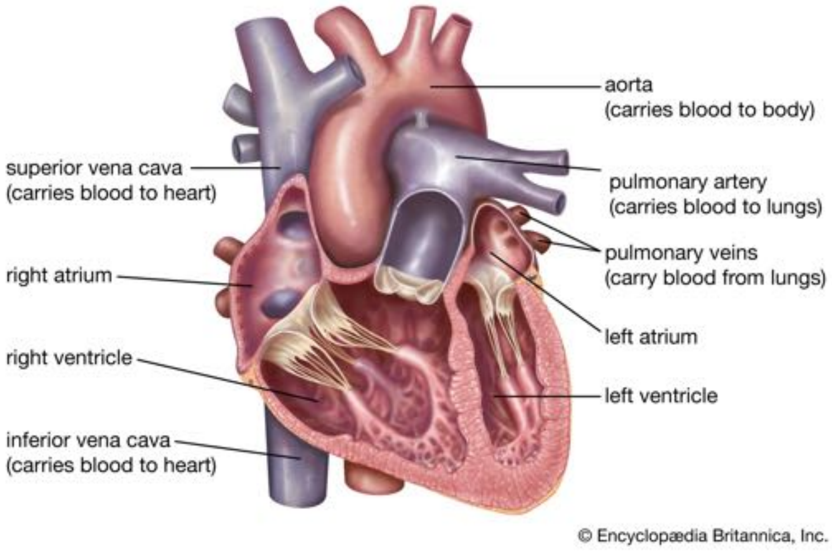

Figure 1: A heart diagram

output a feature map and feed it as an input to the next layer. However, the feature maps are more computationally expensive or contain more parameters than the original images. The job of the pooling layers is to down-sampling the dimensions of the feature map which reduces the number of computations while keeping its important features and using it in the later layers. The dense layers and loss function calculation will be followed after that. In the case of medical images, it is a time-consuming and labor-demanding task to generate pixel-wise labels. It is desirable to develop a weakly segmentation model that only uses a weak rather than a ground-truth label to train the segmentation model. That means the pixel size of the weak label is smaller than the ground-truth label. In the medical imaging community, U-Net is one of the popular models [6] and will be served as the base segmentation model and will be discussed in Section 2.

I will be continuing the work of James Ko, a honors student that is graduated and his work is primarily focused on weakly supervised segmentation which will be discussed in section 2.12. He used U-Net and TernausNet as his segmentation base model and trained four different types of segmentation models with different constraints. A fully-supervised model trained on the ground-truth masks, a weakly-supervised model trained on weak labels without constraint, a weakly-supervised model trained with a size constraint only, and a weakly-supervised model trained with both size and shape constraint. He also used an auto-encoder for enforcing the shape constraints; it will be further discussed in Section 2. In conclusion, the segmentation model that incorporates size and shape constraints trained with the weak labels has the best performance and it is very close to the performance of the model that was trained with fully labeled segmentation masks.

In James' work, there are two stages in his work still deploying information from the true masks. First, he generated and utilized the weak label by applying binary erosion to the ground-truth label. However, in the real application, we are not given with any ground-truth label. Second, in the auto-encoder training process, it uses weak labels to learn the shape embedding of the left ventricle. In this paper, we describe the process to generate these partial masks from the heart MRI scans and training the auto-encoder without using any ground-truth mask. Lastly, we describe a new way to encode the shape constraint to guide the model to segment the left ventricle from a heart MRI.

The main contributions of this paper are the following:

- We generate the synthetic masks through some of our prior knowledge of the shape of the left ventricle and it is implemented using skimage package. They are the training masks that are passing into the auto-encoder.

- We develop an iterative algorithm that is able to generate the partial masks. It consists of finding the center slice, segment and locate the left ventricle, and propagate the location information to the neighbor slices, and repeat the above procedure. It is implemented using skimage package.

- In the new deep learning model pipeline, we revise the model to a weakly supervised manner and incorporate the new shape constraint and universal size constraint in the loss function calculation which guides the model to segment the images and revised the structure of the auto-encoder.

The remaining sections of this paper are as follows: section 2 will go over the works that inspired and contributed to my work and their summaries. Section 3 provides the detailed procedure of the generation of the synthetic masks and the re-work of the shape constraint. Section 4 will go over the results of the experiments that I did this semester. Section 5 will going over the conclusion.

2 Background

2.1 Constrained-CNN losses for weakly supervised segmentation

Constrained Optimization has been largely avoided in deep networks. In the work of [5]Pathak, he addresses constrained deep CNNs in weakly segmentation. The main idea of his work is to model the proposals through a latent distribution. Then, they minimize a KL divergence, encouraging the softmax output of the CNN to match the latent distribution as closely as possible. They impose constrained on the latent distribution instead of the neural network. They introduce a differentiable term, which enforces inequality constraint directly in the loss function, avoiding expensive Lagrangian dual iterates and proposal generation. In addition to the cross-entropy loss function, they added a differentiable term $\mathcal{C}(V_s)$. The new loss function is: $\mathcal{H}(S) + \lambda\mathcal{C}(V_s)$. $\mathcal{H}(S)$ is the cross-entropy loss function, λ is a positive constant, and $\mathcal{C}(V_s)$ enforces the target region constraint.

$$\mathcal{C}(V_s) = \begin{cases} (V_S - a)^2 & \text{if } V_s \leq a \\ (V_S - b)^2 & \text{if } V_s \geq b \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

2.2 Deep learning for cardiac image segmentation: a review

This [2]paper talks about the basic overview of deep learning in the context of medical imaging. It went over the popular deep network architectures, different network layers, evaluation metrics, public data sets, segmentation architectures, and enhancement mechanisms. It is a good paper for me to learn all the basics tool that allows to implement this project.

2.3 Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation

This [3]paper went over the architecture of U-Net. The U-Net compose of an encoder and a decoder. The encoder follows the pattern of convolution and pooling operation. The decoder follows the pattern of upsampling of the feature map and convolution. They particularly used VGG-11 and replaced the fully connected layers with a single convolution layer of 512 channels. The decoder is constructed by using the transposed convolution layer. The number of downsampling and upsampling are the same, to preserve the dimension of the input and output the same. They trained their network on the Inria Aerial Image labeling dataset. The evaluation metric they used is the Jaccard index. The loss function they used is $L = H - \log(J)$. H is the cross-entropy loss and J between masks and corresponding predictions. They also tried out 3 U-Nets with different weights initialization. For the first initialization, they sample from a uniform distribution. The second initialization they did they utilized the VGG-11 pre-trained encoder on the imageNet, and the weights of the decoder are initialized by LeCun uniform initializer. The third thing they did is use the fully pre-trained U-net on the imageNet. The result shows that the pre-trained encoder performs better than the encoder that is not pre-trained.

2.4 Deep learning for atrial segmentation from late gadolinium-enhanced MRIs

This paper is about atrial segmentation.The problem they are facing is the class imbalance problem, left atrial is small and the background is huge. Since the LA is only 0.7% of the volume of the heart, it creates an obstacle to the model to identity LA due to the noisy background. To address this problem, the author used two consecutive networks to segment the left atrial(LA). The first network was used to identify the center of the LA, and crop all the unnecessary background around the center of the LA. Then, they passed this cropped version of the MRI to another neural network. The job of this CNN is to segment the LA. The second problem that faced is the inconsistency in the sizes of the LA anatomical structures. People are tried to use pyramid pooling to solve this. However, due to the cropping function in the first CNN, the pyramid pooling might only show limited benefit from context learning. People also tried dilated convolution layers at the deepest level of their network. This also doesn't improve the result, and waste of memory. The author concluded that the current leading approach for LA segmentation is two-stage CNN.

2.5 Deepusps: Deep robust unsupervised saliency prediction with self-supervision

In this [4]work, the author proposed an unsupervised algorithm that takes different handcrafted methods to generate coarse pseudo-labels, then a deep network(inter-images consistency) takes training images and pseudo-labels to generate consistent label outputs. The label outputs are further refined through self-supervision. Lastly, the refined labels from different handcrafted methods are fused for training the saliency prediction network. In other words, the poor-quality pseudo-labels are further processed through a series of refinement with deep network training, and in the final stage, the refined pseudo-label sets are fused by training a network to minimize the averaged loss between different methods.

2.6 Constrained convolutional neural networks for weakly supervised segmentation

The [5]author proposed a constrained CNN and was able to incorporate arbitrary linear constraints. Since the objective function subject to the linear constraint is not convex, it is hard to directly optimize it. Instead, the author proposed to optimize the latent probability distribution over the semantic label X, and if the network parameters are fixed, then the problem is convex and it is much easier to optimize. The constraints they have are suppression, foreground, background, and size constraint. The way they trained their model is for a fixed latent distribution, for each convnet output they calculated a latent distribution probability distribution $P^{(t)}$ as the closest point in the constrained region and then update the convnet parameters using SGD, and repeat this process many times.

2.7 U-net: Convolutional networks for biomedical image segmentation

In this [6]paper, the authors used U-Net as the segmentation model to segment the neuronal structures in electron microscopic stacks. The U-Net they used consists of the contradicting and expansive path. The contradicting path followed the typical convolution pattern, and the expansive path consists of an upsampling of the feature map followed by 2x2 up-convolution. They also utilized data argumentation with elastic deformation and because of the amount of available data is limited.

2.8 On regularized losses for weakly-supervised CNN segmentation

$$R_{KC}(S) = \sum_k S^{k'} W(1 - S^k) + \gamma \sum_k \frac{S^{k'} \hat{W}(1 - S^k)}{d' S^k} \tag{2}$$

W here can be dense or sparse matrix.

This paper introduces a lot of regularizers from shallow segmentation as loss function. For example, Markov random fild(CRF), conditional random field, and Kernel cut loss. Kernel cut loss combines MRF/CRF loss together, and its loss's gradient descent can be efficiently calculated. In their experiment the best weakly supervised segmentation is achieved with kernel cut loss.

Approximate alternating direction(ADM) method for optimization. Instead of optimizing directly regularized loss with respect to network parameters, proposal methods splits the optimization problem into two easier sub-problems. Replacing the network outputs S_p in the regularization by latent distribution X_p, which is (Kullback-Leibler) KL divergence in this case.

2.9 Unsupervised semantic segmentation by contrasting object mask proposal

This [7]paper, went over incorporating contrastive learning in segmentation. Mask contrast learns pixel embeddings for unsupervised semantic segmentation. The model will try to group similar features together and push away the features that are not similar. If a pair of pixels belong to the same mask, they assume that they should be grouped together.

The objective is to learn a pixel embedding function Ψ_θ parameterized by a neural network with eights θ, that maps each pixel i in an image to a point z, on a D-dimensional normalized hyper-sphere.

Learning Image-Level Representations

The positive and negative views are used to learn the image-level representation. A positive view means that both images contain the same object. A negative view means that both images don't contain the same object. They did this by training an image embedding function Ψ to maximize the agreement between positive pairs and minimize the agreement between the negative pairs.

Learning Pixel-Level Representations

If two pixels are from the same object, then it should maximize the agreement between their pixel embedding. This is what is called the pull force. If the pixels are from a dissimilar object, then they should be mapped further apart. The pixel embedding function is going to maximize the agreement between pixels and an augmented view of the object they belong to and minimize the agreement with other objects. In summary, their optimization objective is optimizing the alignment of pixel embeddings based on shared pixel ownership.

2.10 contrastive learning of global and local features for medical image segmentation with limited annotations

Similiarly, this [1]paper also went over the contrastive learning in the context of segmentation of medical images.

Global contrastive loss

minimizing the loss increases the similarity between the representations while increasing the dissimilarity between the representation of x and those of dissimilar images. Also, they used cosine similarity between two vectors to measure the similarity in the representation space. In the global contrastive loss. They hope their model is able to distinguish and learn group ith group together and jth group together, etc. They also integrate domain and problem-specific information to improve the effectiveness of the resulting self-supervised learning process.

Local contrastive loss

What they here assume is the s_i group of volume 1 and volume 2 should be similar. The architecture consists of the encoder, decoder, and convolutional layer. The decoder blocks are trained using a local contrastive loss. They expected that different local regions within the l groups to be dissimilar, while each local region in the l group remains similar across intensity transformations.

2.11 Positional contrastive learning for volumetric medical image segmentation

In this [9]paper, they addressed the issue of many false negative pairs that result in degraded segmentation quality. This is because the same partition for different volumes is considered positive, and a different partition is considered negative. However, the last few slices of one partition looked very similar to the next beginning of a few slices of the next partition. This leads to a false negative. They addressed this issue by leveraging the domain-specific cue of medical images, and relative position. Slices that are close are considered positive pairs while those that are far apart are considered negative.

The architecture of their model is as followed the 2D images are taken from the xy-plane, and each of them has its z-value(relative position). If the change in position of two images is less than 1, then it is a positive pair. If the change in position of the two images is greater or equal to 1, then it is a negative pair. They used serval images to train the U-Net's encoder and used the trained encoder as the initialization in the fine-tuning stage.

2.12 Constraint Weakly Segmentation Model with Cardiac MRI

This paper purposes and trains five different segmentation models which use different constraint on the loss function calculation inspired by [5]. A fully-supervised model trained on the ground-truth masks, a weakly supervised model trained on weak labels without constraint, a weakly-supervised model trained with a size constraint only, and a weakly-supervised model trained with both size and shape constraint. These models are elavuated with dice and Intersection of Union(IoU) scores which is summarized in table 1. The best model shown in table 1 is the U-Net trained with size + shape constraints.

Let N denotes the number of training images and P denotes the number of pixels per image. Let $X_i \in \mathbb{R}^P$ denotes the ith training image, $Y_i \in \{0,1\}^P$ denote the corresponding ground-truth mask, and $S_i \in [0,1]^P$ denote the ith predicted mask as a vector of P probabilities

For the full supervision, the loss function is binary cross entropy as shown in equation 3, and the full loss function is as shown in equation 4. We are penalizing all the predicted pixels are not matching with the truth mask which is just pixel-wise prediction.

$$BCE(s, y) = -\sum_{p=1}^{P} y_p \cdot \log s_p + (1 - y_p) \cdot \log 1 - s_p \tag{3}$$

$$L_{full} = -\sum_{i=1}^{N} BCE(S_i, Y_i) \tag{4}$$

For the weak supervision, since we are only using the weak masks, we are only penalizing the false positive pixels because we don't know the location of the true negative pixels. The loss function for weak supervision is shown in equation 5, and the full weak loss function is shown in equation 6.

$$\text{PartialBCE}(s, \omega) = -\sum_{\omega=1} \log s_p \tag{5}$$

$$L_{weak} = -\sum_{i=1}^{N} \text{PartialBCE}(S_i, W_i) \tag{6}$$

For the weak supervision with size constraint, we are penalizing the model if the pixels size of the predicted mask is smaller or bigger than a user defined min l_i and max μ_i threshold for ith image. The equation for the size constraint is shown in equation **??**, and the full loss function for size constraint is shown in equation 15.

$$L_{\text{size}} = \sum_{i=1}^{N} \begin{cases} (l_i - |S_i|)^2 & \text{if} |S_i| < l_i \\ (|S_i| - \mu_i)^2 & \text{if} |S_i| > \mu_i \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Model	Strategy	Description	λ	α	Dice	IoU
U-Net	full	Full supervision			94.162	89.135
	weak-size	Size constraint only	$3 \cdot 10^{-4}$		85.418	75.162
	weak-ss	Size + shape constraints	$3 \cdot 10^{-4}$	$3 \cdot 10^{-3}$	86.883	77.412
TernausNet	full	Full supervision			94.454	89.644
	weak-size	Size constraint only	10^{-3}		86.228	76.377
	weak-ss	Size + shape constraints	10^{-3}	$3 \cdot 10^{-3}$	86.810	77.281

Table 1: Result for different strategy, model, parameter, metrics.

$$L_{\text{weak-size}} = L_{\text{weak}} + L_{\text{size}} \tag{8}$$

l_i and μ_i are the individual size bound that is generated from the ground-truth mask which too much information leaks into the weak segmentation model.

The weak supervision with shape constraint consists of two training stages. In the first stage, we have to train a denoising auto-encoder using the weak masks. In the next stage, we have to train segmentation model and use the encoder part of the denoising auto-encoder as a verification tool to enforce the shape constraint of the predicted masks. The guidance is as followed: we are penalizing the model if the distance of the encoded shape embedding of predicted mask is far away from the distance of the shape embedding of the weak mask. In other words, we are penalizing the model if the shape of predicted mask is completely different than the shape of weak mask.

$$L_{\text{shape}} = \sum_{i=1}^{N} \max\left(||E(S_i) - E(W_i)||_2 - \epsilon, 0\right) \tag{9}$$

$$L_{\text{weak-shape}} = L_{\text{weak}} + \alpha \cdot L_{\text{shape}} \tag{10}$$

Let E be the encoder from the de-noising auto-encoder, and $E : \mathbb{R}^P \to \mathbb{R}^Q$, and ϵ is the 99th percentile of trained shape latent vectors.

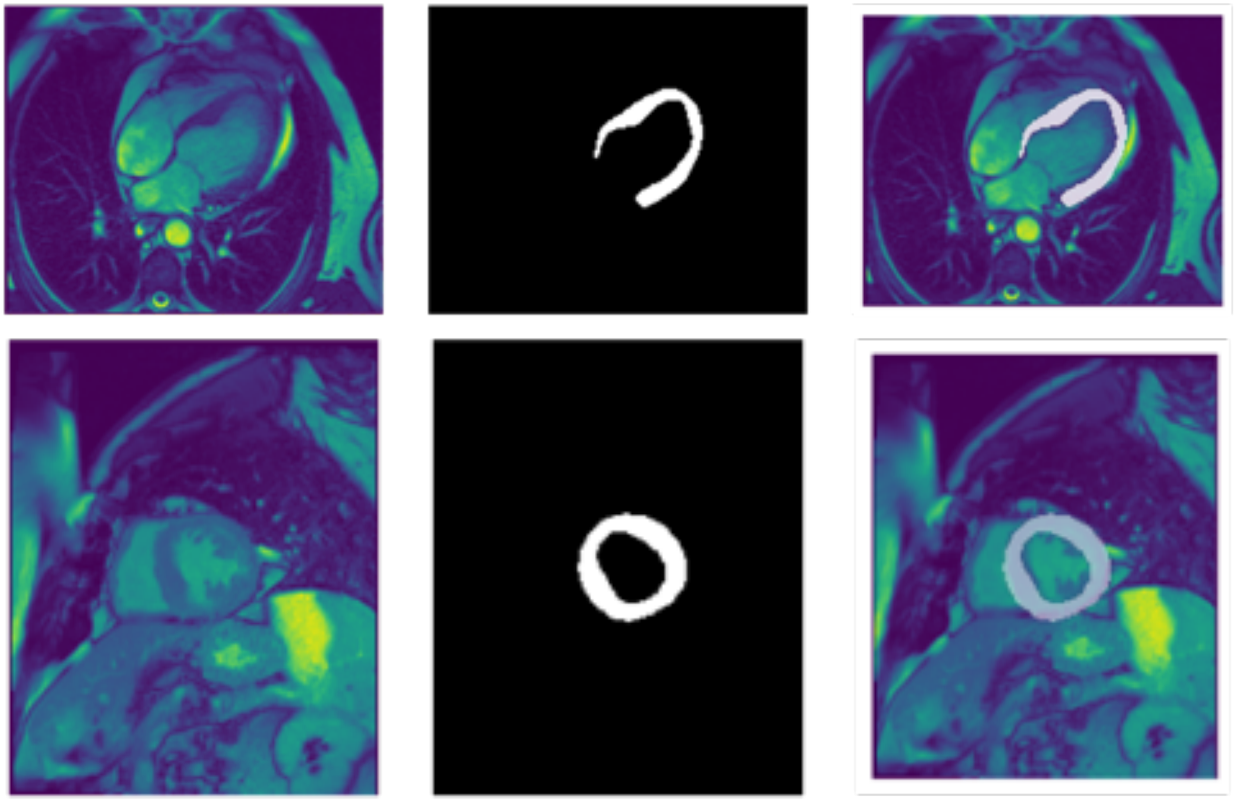

Figure 2: The top row is the long-axis view of the left ventricle, and the bottom row is the short-axis view of the left ventricle. Start from left to right is showing the left ventricle, mask, and left ventricle annotated with the mask.

2.12.1 Weak Mask and Weak Size Generation

All the results and analysis above are produced by the model that is trained with the weak masks. The procedure of generating the weak masks is applying the binary erosion to every ground-truth mask with the kernel size of 3 × 3, and the individual size bounds for each weak mask are obtained by multiplying 0.9 and 1.1 by the number of pixels.

3 Dataset

The dataset that I will be using is a publicly available dataset left ventricle segmentation challenge (LVSC). This dataset contains 21918 cardiac image slices from 100 unique patients in a time series format and was chosen because it has a large amount of data that contains both short and long-axis view images. Each image is uniquely identified by (ptid, slice, phase). The phase number ranges from 0 to 34 which describes the time step and the time evolution of the heart. The slice number indicates the location of the heart and it varies for different patients ranging from 1(leftmost) to 23 (rightmost). Ptid is the patient ID which will be uniquely assigned to each patient. I will be utilizing the short-axis images because they encode a more detailed description of the left ventricle. Some examples from the heart MRI scan and ground-truth masks from the LVSC dataset are shown in figure 2.

4 Methodology

Since the methodology described in section 2.12 is not fully weakly supervised, namely the training process of the auto-encoder and segmentation model involves using the weak masks and they are generated from the ground-truth masks. Our goal is to revise the model, so that model is trained only with the weak masks generated from the raw heart images. Furthermore, the shape constraint loss function equation 9 defined in section 2.12 could be better improved which will be discussed below. Our experimental models consist of two relevant models that are adapted from section 2.12:

- a weakly-supervised model trained with a shape constraint only (weak-shape).
- a weakly-supervised model trained with both shape and size constraint (weak-ss).

The training pipeline for weak-shape and weak-ss are shown in figure 3. The loss function for shape constraint is described in equation 9 and size constraint is described in equation 7.

4.1 Synthetic Masks

In order to enforce the shape generated by the segmentation model, we are utilizing a trained encoder from a de-noising auto-encoder which is shown in figure 4. In the training process, the de-nosing auto-encoder takes in the true masks and tries to learn the features

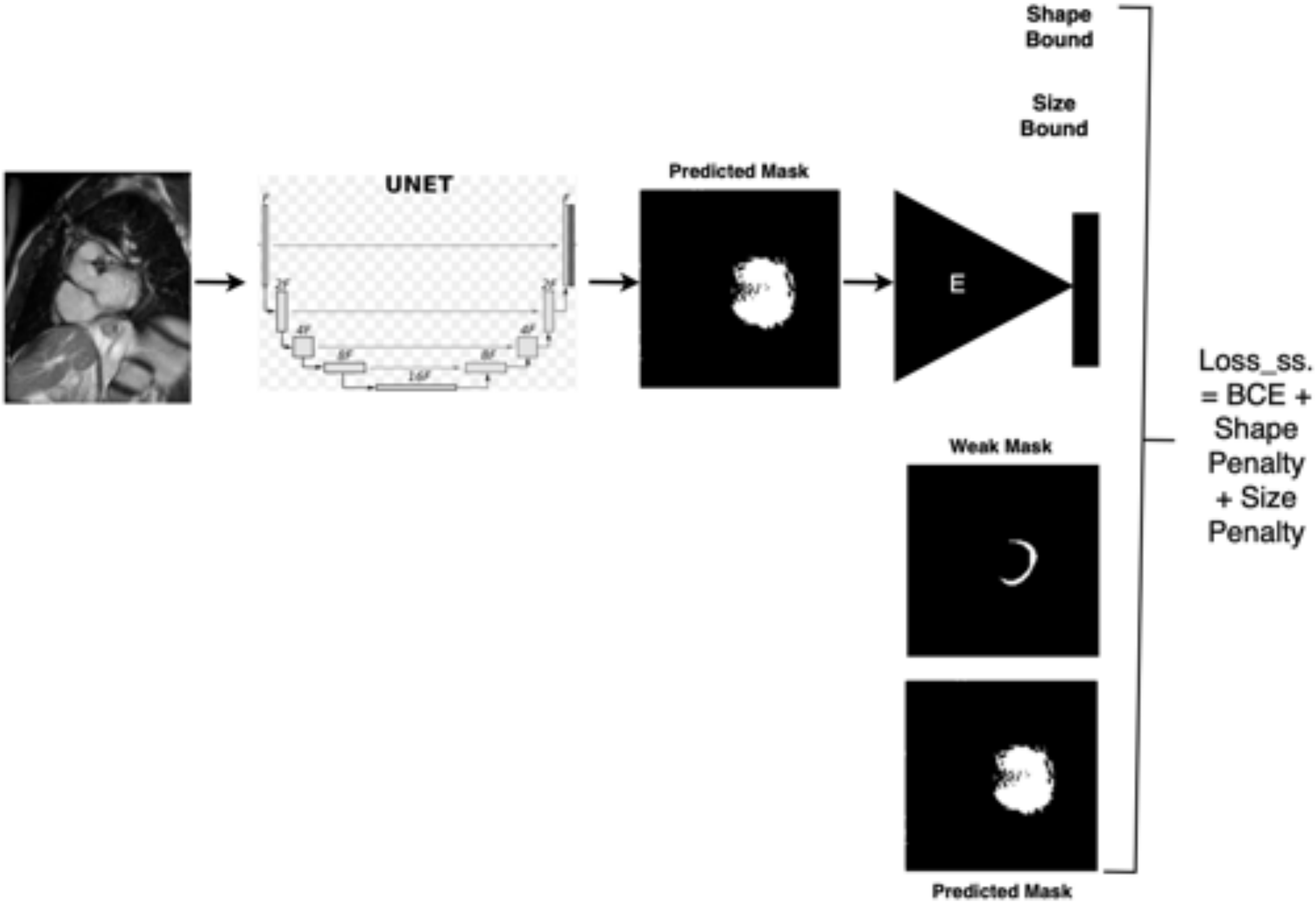

Figure 3: Pipeline for weakly segmentation model with shape and size constraint

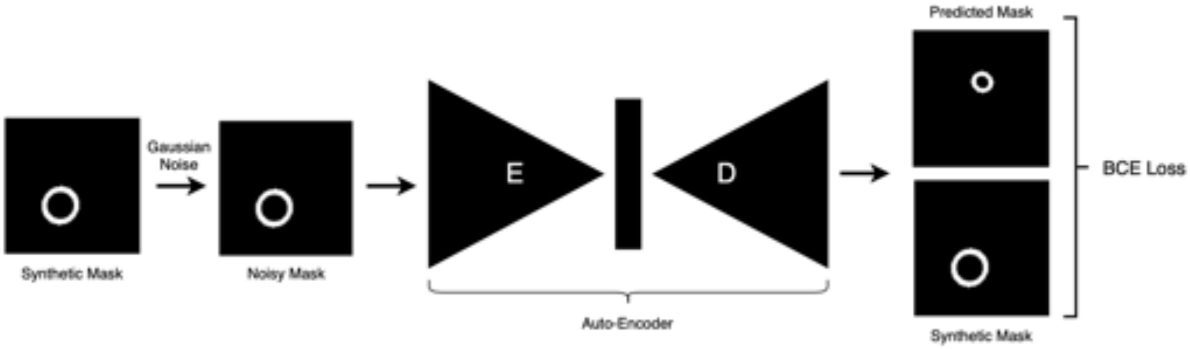

Figure 4: Pipeline for a denosing auto-encoder

of the true masks. However, the true masks are like an irregular-elliptical donuts, with a hollowed center. With this prior knowledge in mind, we can generate artifactual/synthetic masks that have a similar shape compared with the true masks and this decreases the information that we seeking from the true mask.

The job of the synthetic masks is to guide our autoencoder to learn the features of the segmented left ventricle. The synthetic masks are generated by using Python's library skimage. The procedure to generate the synthetic masks are as follows: randomly choose the radius and center of the two axes of an ellipse, it has to be within some reasonable range, and randomly rotate the ellipse. The above procedure will be repeated once or twice with an equal probability. This is because some masks don't have a concentric circle and some do. If two ellipses are generated, then the synthetic mask will be a bigger radius ellipse minus the smaller radius ellipse. We can generate as many of them as we want to achieve the goal of learning the features. Some ground-truth and synthetic masks are shown in the figure 5.

4.2 Shape Constraint

In section 2.12, the pipeline of his model with only the shape constraint is as follows: passing the training MRI images to the segmentation model, outputing a predicted mask, and passing that predicted mask into the trained encoder. The output of this encoder will be an embedding vector that describes the shape of the predicted mask. Based on this

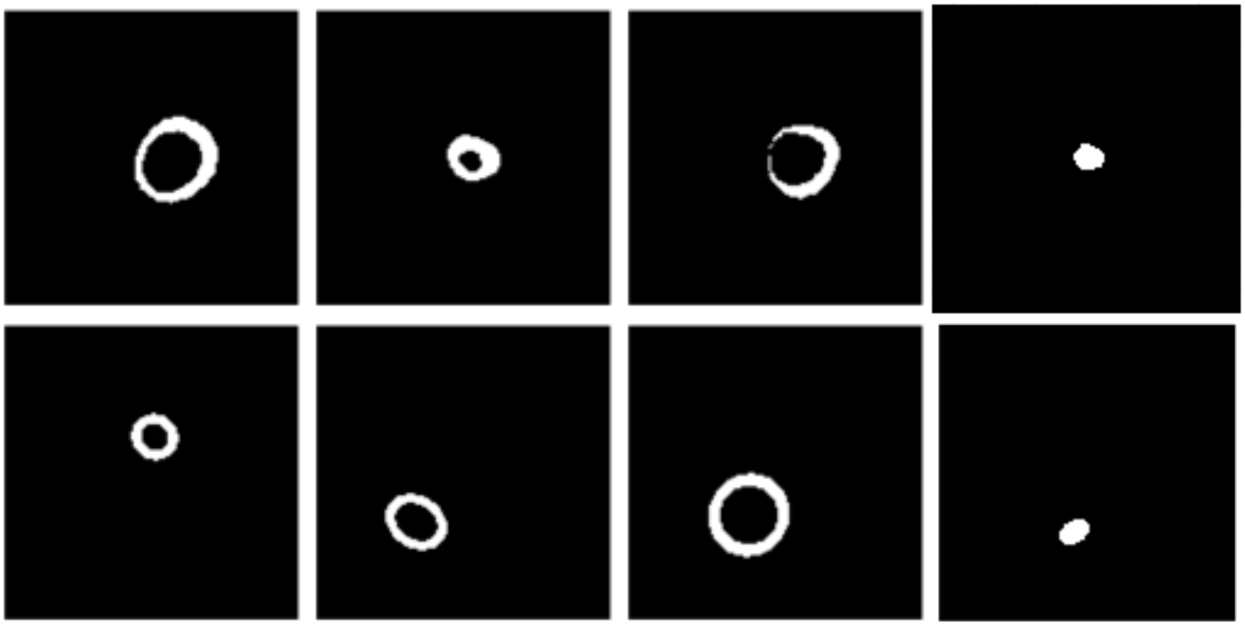

Figure 5: The top row is the ground-truth labels for the left ventricle segmentation, and the bottom row is the synthetic masks for the left ventricle segmentation.

embedding vector, he introduces the shape constraint term in the loss function which will encourage the shape of the predicted mask and weak label to be similar which is shown in equation 9. This equation penalizes the model if the L2 norm of the difference between the encoded predicted mask and the encoded noisy mask is greater than ϵ which doesn't fully enforce the shape of the two masks to be similar.

Let N denotes the number of synthetic masks, E denotes the trained encoder which $E : \mathbb{R}^P \rightarrow \mathbb{R}^Q$, P denotes the number of pixels in the synthetic mask, Q denotes the dimension of the embedding vector and will be further discussed in section 5, SM$_{avg}$ denotes the average of all the shape embedding vectors that live in the valid L2 norm ball, S_i be the ith predicted mask made by the segmentation model, and ϵ denotes the shape distance bound. The improved shape constraint is shown in equation 11, and the full loss equation is shown in equation 12.

$$L_{\text{shape}} = \sum_{i=1}^{N} \max\left(||E(S_i) - SM_{avg}||_2 - \epsilon, 0\right) \tag{11}$$

$$L_{\text{weak-shape}} = L_{\text{weak}} + \alpha \cdot L_{\text{shape}} \tag{12}$$

This is implemented by extracting the latent vectors from the embedding space, and in the embedding space, we are expecting that all the reasonable features will be in an L2 norm ball. The center (SM_{avg}) of that l2 norm ball will be calculated by finding the mean of all the latent vectors as shown in equation 13, and the radius of the l2 norm ball will be determined by the shape distance bound (ϵ).

$$SM_{avg} = \frac{1}{N}\sum_{i=1}^{N} E(SM_i) \tag{13}$$

The distance bound (ϵ) is calculated by taking the 99th percentile of L2 distance of all pairs of latent vectors and center, and we plotted them as shown in figure 6. By doing this, we are discarding the outliers or the unreasonable latent vectors that are very far from the center. We can use this distance bound as a threshold. If the L2 norm of the encoded predicted mask is greater than ϵ, then penalize it, and not penalize if it is less than ϵ which describes by equation 11. In other words, if the shape of the predicted mask is not left ventricle alike, then its L2 is expected to be greater than ϵ, and versa visa. This new shape constraint enforces all the L2 norm of the predicted masks to be less than ϵ and lives in that L2 norm hyper-sphere defined by the center (SM_{avg}) and radius (ϵ).

4.3 Universal Size Constraint

The weakly segmentation model with size constraint alone has a promising result shown in table 3, and the equation for the size constraint is shown in equation 7. However, each l_i and u_i are obtained by multiplying 0.9 and 1.1 by the number of pixels from the ith true mask which uses too much information from the true mask. Our goal is somehow to obtain l_i and μ_i without any information from the true mask. This information could be obtained from domain knowledge. However, we have an individual size bound for each slice number

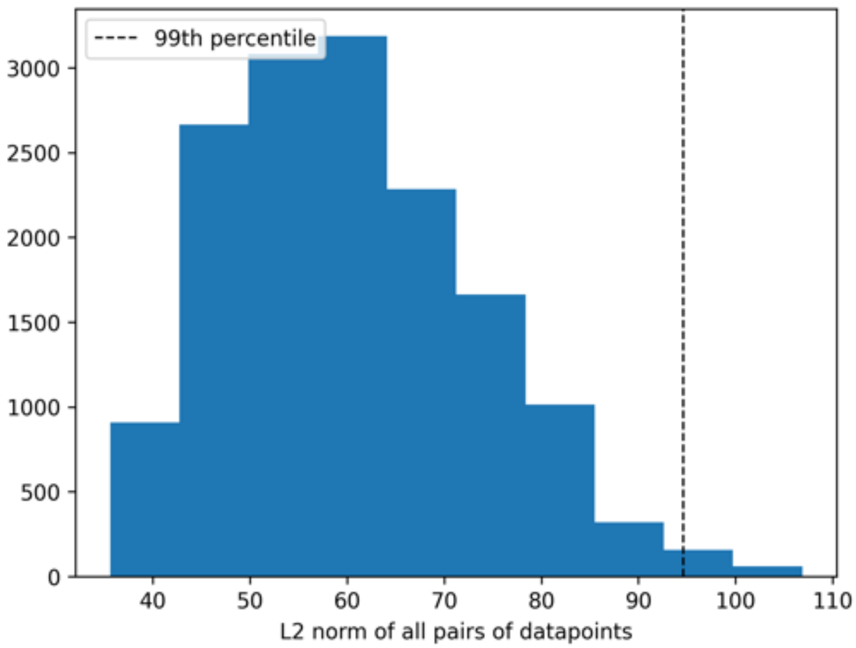

Figure 6: This histogram shows the l2 norm distance of all pairs of the center of the l2 norm ball and the data points. The dashed line indicates the 99th percentile distance and the remaining 1% is to the left of the dashed line.

and time step, we would rather have a universal upper and lower size bound for all the images. The size bound is obtained from [8]. It talks about the size statistics of various components of the heart. For the Left ventricle was thickest in the basal septum with a mean thickness of 8.3 mm and 7.2mm and thinnest in the midventricular anterior wall with 5.6 mm and 4.5 mm for men and women, respectively. This information is being used in the size constraint calculation. The universal size constraint equation is defined in equation 14. Let l_{univ} and μ_{univ} denote the universal size lower and upper bound, respectively.

$$L_{\text{usize}} = \sum_{i=1}^{N} \begin{cases} (l_{\text{univ}} - |S_i|)^2 & \text{if} |S_i| < l_{\text{univ}} \\ (|S_i| - \mu_{\text{univ}})^2 & \text{if} |S_i| > \mu_{\text{univ}} \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

$$L_{\text{weak-usize}} = L_{\text{weak}} + L_{\text{usize}} \tag{15}$$

4.4 Universal Size and Shape Constraint

We could combine two constraints together which is discussed in section 2.12 to form a stronger constraint, namely the universal size and the new shape constraint. This combined constraint guides the model to produce a predicted mask that has the compatible size and shape of the left ventricle. The equation for the combined constraint is shown in equation 16.

$$L_{\text{weak-sus}} = L_{\text{weak}} + L_{\text{usize}} + \alpha \cdot L_{\text{shape}} \tag{16}$$

L_{weak} penalizes the model for the false positive pixels. L_{usize} penalizes the model if the size of the predicted mask is outside the bound of l_{univ} and μ_{univ}. α is the shape coefficient which is one of the hyper-parameters. L_{shape} penalizes the model if the shape of the predicted mask is not like the left ventricle.

4.5 Weak Mask Generation

The generation of the weak masks consists of many stages and it is described by the pseudocode shown in algorithm 1, and it is implemented by using Python's skimage package. The format of the dataset is described in section 3. We assumed that the location of the left ventricle will not be changing that much from time step 0 to 23, therefore, within the same slice number, all the time phases for that specific slice number will share the same coordinate. However, the location of the left ventricle will vary a lot between different slice numbers. Therefore, the procedure of locating the left ventricle between different slices will be much different than between time steps.

First, we want to find the center slice (Slice_{c}) of a particular patient because it is the easiest to segment the left ventricle since it has the largest area compared with all other slices. We know that the location of the left ventricle must exist somewhere around the center area of the heart MRI scan. We crop the image from an index of 45 to 90 for both axes and reduce the image size 128 × 128 to 50 × 50 to get the center area of the image. This reduces the amount of noise or errors in locating the left ventricle. Next, using histogram equalization to increase the global contrast and Canny edge detection to segment the edges of the image and get rid of unnecessary information. Finally, using the Hough circle transformation to detect the location (c_x and c_y) of the left ventricle. Based on c_x and c_y, we can continue to locate the left ventricle for $\mathrm{Slice_{c-1}}$ and $\mathrm{Slice_{c+1}}$ using the above algorithm to get four new coordinates (c_{x+1},c_{x-1},c_{y+1},c_{y-1}). The coordinate of the center slices is shared between $\mathrm{Slice_{c+1}}$ and $\mathrm{Slice_{c-1}}$, then the coordinate of $\mathrm{Slice_{c+1}}$ propagate to $\mathrm{Slice_{c+2}}$ and so on, also the same for $\mathrm{Slice_{c-1}}$. Furthermore, we shrink the image dimension of 50 × 50 to a smaller window based on the radius returned by the hough circle transformation which deletes more noisy data; the smaller the radius the smaller the window will be, and versa visa. Lastly, based on the location information we found above, we could define two circles. The inner circle is the left ventricle and the outer circle is the myocardium and left ventricle. However, we only want the myocardium (myo) which is outer minus the inner circle. Sometimes, myo is not perfect and could be improved by only taking the pixel values that are in the 2% percentile of myo's pixel intensity.

5 Results

In section 3, we discussed having an encoder (E) to enforce the shape constraint, $E : \mathbb{R}^P \rightarrow \mathbb{R}^Q$. The dimension of Q has a tremendous impact on the performance of the segmentation model which is summarized in table 2. We can see that when the dimension of the latent vector is 2048, the segmentation model with only shape constraint reaches the dice score of 0.478 on the testing set. Whereas, when the dimension of the latent vector is 512 or 1024, the dice score is 0.067 or 0.033, respectively. This suggests that 512 or 1024 is not big enough to encode all the information that is needed for describing the shape feature. On the other hand, 2048 is the right dimension for the latent vector which will be the

Algorithm 1 Partial Masks Generation

```
$N \leftarrow 100$
$k \leftarrow$ number_slices
$p \leftarrow 0$
masks $\leftarrow$ []
for patient $\leftarrow$ 1 to $N$ do
    center_slice $\leftarrow$ find_center_slice(patient)
    image $\leftarrow$ get_image(patient, center_slice)
    center_coordinate $\leftarrow$ find_center_coordinate(patient, image)
    coordinate $\leftarrow$ []
    for $i \leftarrow$ center_slice-1 to $k$ do
        image $\leftarrow$ get_image(patient, i)
        if $i$ == center_slice-1 then
            coordinate.append(find_coordinate(patient, center_coordinate, image))
        end if
        coordinate.append(find_coordinate(patient, coordinate[i-1], image))
    end for
    for $j \leftarrow$ center_slice+1 to $p$ do
        image $\leftarrow$ get_image(patient, j)
        if $j$ == center_slice+1 then
            coordinate.append(find_coordinate(patient, center_coordinate, image))
        end if
        coordinate.append(find_coordinate(patient, coordinate[j+1], image))
    end for
    for $i \leftarrow 1$ to $k$ do
        image $\leftarrow$ get_image(patient, i)
        circle $\leftarrow$ hough_circle(canny_edge_detect(hist_equalization(image)), coordinate[i])
        mask.append(gen_mask(circle, coordinate[i]))
    end for
end for
return masks
```

dimension used in the auto-encoder for later experiments. For demonstration purposes, we also reconstructed some of the masks which will be discussed in the appendix section.

From table 3, we can see the segmentation model with only shape/shape+universal size constraint trained with weak masks have a dice score of 0.66 and 0.81, respectively. This shows that the universal size constraint didn't better perform than the size constraint(section 2.12) and we can see this difference in table 3. This is expected because the individual size bound is stronger than the universal size bound; the purpose of the universal size bound is to make sure the predicted mask is within some reasonable range. On the other hand, the new shape constraint indeed improves the performance of the model by a significant amount (0.66 to 0.81). Whereas, the old shape constraint didn't improve the performance of the model significantly (0.85 to 0.87) which is shown in table 1. This suggests the new shape constraint (equation 11) outperforms the old shape constraint (equation 9). The results above prove that the universal size + new shape constraint model works well. However, the performance of the model decreases after we switch to the partial masks (partial masks_v1, partial masks_intersect, partial masks_v2) which are generated from raw heart MRI.

Theoretically speaking, if we are able to generate partial masks like weak masks, then we can achieve a similar dice score. Partial Masks_v1 is the first version of the partial masks that we generated from the raw heart images which are able to reach a dice score of 0.67. Partial Masks_intersect is generated by taking the intersection of Partial Masks_v1 and true masks which yield a better dice score of 0.77. This suggests that the low dice score is caused by the false positive pixels and could be improved by having more true positive pixels. This leads to our hypothesis that if we increase the true positive and decrease the false positive pixel of the partial masks, then it should able to achieve a better dice score. Partial Masks_v2 which are generated based on this idea, the number of pixels per image is around 3 to 4, and the implementation was described in section 4.4. However, the dice score of partial masks_v2 has a dice score of 0.5. From all the experiments and observations above, this suggests that in order to achieve a high segmentation dice score, the partial masks must

Dimension of the latent vector	λ	α	Dice
2048	0.1	0.001	0.478
1024	0.1	0.001	0.033
512	0.1	0.001	0.067

Table 2: Investigating the dimension of the latent vector and its impact on the segmentation. The above results are obtained by using U-Net as the base model trained with weak masks, auto-encoder was trained with true masks, and the strategy used is weak supervision with only shape constraint.

Strategy	Description	Type of Mask Used	Dice
weak-us	Univerisal size constraint	Weak Masks	0.66
weak-sus	Univerisal size + Shape constraints	Weak Masks	0.81
weak-us	Univerisal size constraint	Partial Masks_v1	0.54
weak-sus	Univerisal size + Shape constraints	Partial Masks_v1	0.67
weak-us	Univerisal size constraint	Partial Masks_intersect	0.63
weak-sus	Univerisal size + Shape constraints	Partial Masks_intersect	0.77
weak-sus	Univerisal size + Shape constraints	Partial Masks_v2	0.50
weak-size	Size constraints	Weak Masks	0.85
weak-ss	Size + Shape constraints	Weak Masks	0.87
full	Full supervision	Ground-Truth Masks	0.94

Table 3: Segmentation results by using different type of masks and investigate the effect of different partial masks on the segmentation results.

satisfy the following conditions:

1. The true positive pixels $\gg$ false positive pixels for every partial mask image.

2. The number of pixels per image has to be sufficient enough.

The weak supervision is only used in the BCE calculation equation which is shown in equation 5. The job of BCE is to guide the model to locate the left ventricle. The false positive pixels from the partial weaks are misguiding the model to locate the location of the left ventricle. This leads to a low dice score. The partial masks don't have to be perfect and it will be a subset of the true mask.

6 Conclusion

Deep learning models are a powerful tool in the medical imaging community. It could help doctors to identify internal organs, calculate the ejection fraction of the left ventricle, etc. However, it is a time-consuming and labor-demanding task to obtain pixel-wise annotations for masks. As that being said, the segmentation model that doesn't use any annotated masks or use weak masks as the training data will be ideal. The previous work (section 2.12) proves that shape and size constraints with weak supervision can enforce segmentation masks to have the characteristics of the left ventricles and have a comparable segmentation result to the full supervised models. In this paper, we discuss the improvements that we have made to the previous work (section 2.12), the pipeline of semantic segmentation with weak supervision and domain knowledge, and the generation of partial masks from raw heart MRI images. We tested our model on the LVSC dataset and the results suggest that the weakly supervised model that incorporates the universal size and shape constraint is the best model which has a dice of 0.67 and without any information leakage from the true mask.

Here are some direction of continuing and bettering this model:

- **Generate better partial masks.** The performance of the weakly segmentation model is depending on the quality of the partial masks as we can see in table 3. This suggests that generating better-quality of partial masks is straightforward to improve the performance of the model.

- **Use Varational Auto-encoder.** Denoising auto-encoder is not the state-of-art model for representation learning. One thing we could try to do is switch the denoising auto-encoder with the denoising variational auto-encoder.

- **Use other cardiac dataset.** All the experiments are done with the LVSC dataset. It is desirable to test this model on other heart MRI scan datasets and see its performance. For example, the UK biobank dataset,

- **Investigate the size of the partial masks.** The size of the partial masks can be one of the factors that impact the performance of the segmentation. Small size could leads to low dice scores, and high size will have a high dice score and is hard to generate and defeat the purpose of weakly-supervised. As that being said, the size of the partial mask is worth the time to investigate.

7 Acknowledgments

I would like to thank my committee chair Professor Madalina Fiterau and my second committee member Professor Evangelos Kalogerakis for their time in supporting and guiding me through this difficult process. I would also like to thank Ke Xiao for answering and clarifying my questions and setting up meetings with me on a regular basis to discuss the potential solutions to the problems that we are facing and sending me resources that help to learn the materials that I needed for the research.

8 Appendix

Below are some figures 7, 8, and 9 which provided some visualization of reconstructed masks generated by the auto-encoder which served as debugging purposes while conducting the experiment. The last figure is some predicted masks made by the best segmentation (table 3) that has a dice of 0.77.

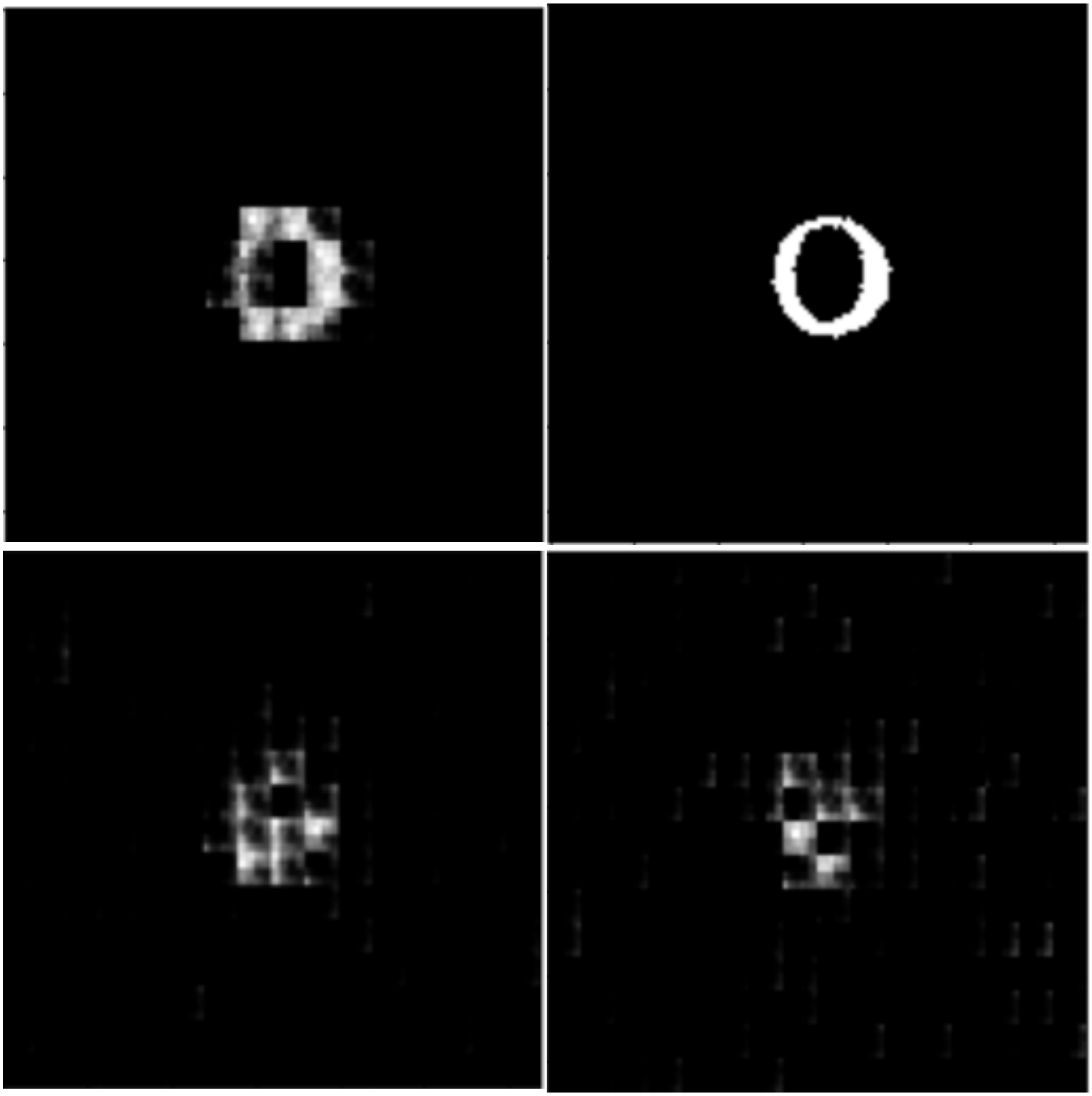

Figure 7: Here are some masks generated by using a latent vector of size 512. The first row and first column is the reconstructed mask after the noise true mask (first row and second column) is passed into the auto-encoder. The second row is the masks generated by inputting a random vector that has an L2 norm of 30.

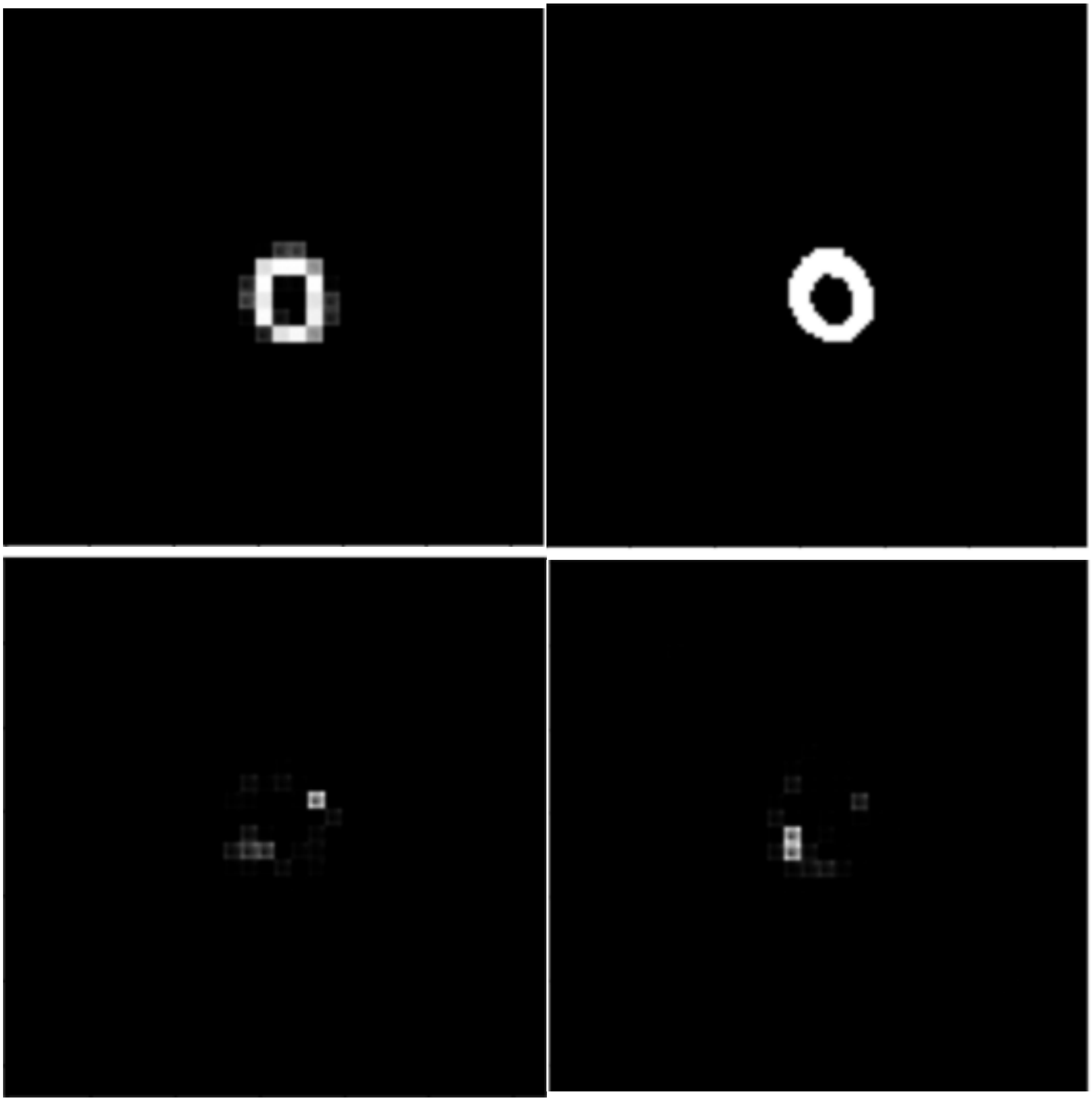

Figure 8: Here are some masks generated by using a latent vector of size 1024. The first row and first column is the reconstructed mask after the noise true mask (first row and second column) is passed into the auto-encoder. The second row is the masks generated by inputting a random vector that has an L2 norm of 30.

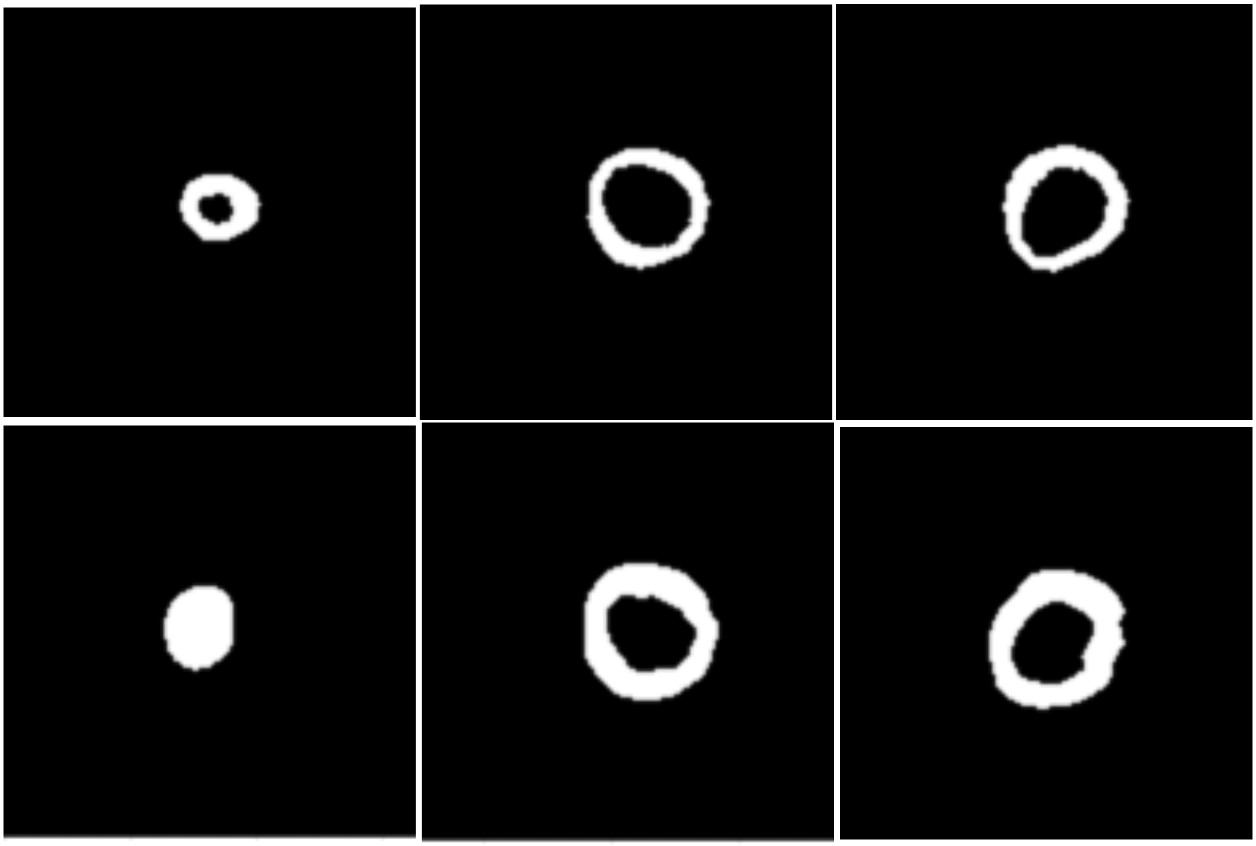

Figure 9: The first row is the true masks, and the second row is the corresponding predicted masks made by the best segmentation model that is trained with Partial Masks_v1 which doesn't use any information from the true mask.

References

[1] Krishna Chaitanya et al. *Contrastive learning of global and local features for medical image segmentation with limited annotations.* 2020. DOI: 10.48550/ARXIV.2006.10511. URL: https://arxiv.org/abs/2006.10511.

[2] Chen Chen et al. "Deep learning for cardiac image segmentation: a review". In: *Frontiers in Cardiovascular Medicine* 7 (2020), p. 25.

[3] Vladimir Iglovikov and Alexey Shvets. "Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation". In: *arXiv preprint arXiv:1801.05746* (2018).

[4] Duc Tam Nguyen et al. "Deepusps: Deep robust unsupervised saliency prediction with self-supervision". In: *arXiv preprint arXiv:1909.13055* (2019).

[5] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. "Constrained convolutional neural networks for weakly supervised segmentation". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1796–1804.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[7] Wouter Van Gansbeke et al. "Unsupervised semantic segmentation by contrasting object mask proposals". In: *arXiv preprint arXiv:2102.06191* (2021).

[8] Jeroen Walpot et al. "Left Ventricular Mid-Diastolic Wall Thickness: Normal Values for Coronary CT Angiography". In: *Radiology: Cardiothoracic Imaging* 1 (Dec. 2019), e190034. DOI: 10.1148/ryct.2019190034.

[9] Dewen Zeng et al. *Positional Contrastive Learning for Volumetric Medical Image Segmentation.* 2021. DOI: 10.48550/ARXIV.2106.09157. URL: https://arxiv.org/abs/2106.09157.