# Few-shot Learning by a Cascaded Framework with Shape-constrained Pseudo Label Assessment for Whole Heart Segmentation

Wenji Wang, Qing Xia, Zhiqiang Hu, Zhennan Yan, Zhuowei Li, Yang Wu, Ning Huang, Yue Gao, Dimitris Metaxas and Shaoting Zhang

***Abstract*— Automatic and accurate 3D cardiac image segmentation plays a crucial role in cardiac disease diagnosis and treatment. Even though CNN based techniques have achieved great success in medical image segmentation, the expensive annotation, large memory consumption, and insufficient generalization ability still pose challenges to their application in clinical practice, especially in the case of 3D segmentation from high-resolution and large-dimension volumetric imaging. In this paper, we propose a few-shot learning framework by combining ideas of semi-supervised learning and self-training for whole heart segmentation and achieve promising accuracy with a Dice score of 0.890 and a Hausdorff distance of 18.539 mm with only four labeled data for training. When more labeled data provided, the model can generalize better across institutions. The key to success lies in the selection and evolution of high-quality pseudo labels in cascaded learning. A shape-constrained network is built to assess the quality of pseudo labels, and the self-training stages with alternative global-local perspectives are employed to improve the pseudo labels. We evaluate our method on the CTA dataset of the MM-WHS 2017 Challenge and a larger multi-center dataset. In the experiments, our method outperforms the state-of-the-art methods significantly and has great generalization ability on the unseen data. We also demonstrate, by a study of two 4D (3D+T) CTA data, the potential of our method to be applied in clinical practice.**

***Index Terms*— whole heart segmentation, pseudo label, quality assessment, self-training, semi-supervised**

## I. Introduction

Cardiovascular disease (CVD) is still the leading global cause of death, and heart disease remains the number one cause of death in the US [1]. A comprehensive analysis of patient-specific cardiac structure and motion is fundamental for understanding cardiac function, early detection, and accurate treatment of CVDs. There are different noninvasive imaging technologies used for understanding and diagnostic purposes in cardiology [2], such as computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound (US). Among these imaging modalities, cardiac CT is fast, low cost, and generally of high quality [3]. Based on the high-resolution CT scan, a 3D model of the whole heart can be built for quantitative analysis. Furthermore, a 4D (3D+T) CT can be used for 3D motion and strain analysis to provide intuitive information about heart function by visualization.

Accurate heart segmentation from an image is a prerequisite for the construction of a 3D or 4D heart model. In the task of whole heart segmentation (WHS), each of the individual heart substructures, including the left ventricle (LV), right ventricle (RV), left atrium (LA), right atrium (RA), the myocardium of LV (MYO), ascending aorta (AO) or the whole aorta, and the pulmonary artery (PA), needs to be extracted from volumetric images [4]. Although manual delineation can provide accurate labels, it requires professional domain knowledge and is laborious and time-consuming due to the large variations in the shape of the heart, low contrast between different substructures, and a lot of 2D slices in a 3D image. It can take hours to label a whole heart [5]. Therefore, automatic solutions for efficient cardiac segmentation are desired. Zhuang et al. [4] provided a benchmark in the Multi-Modality Whole Heart Segmentation Challenge (MM-WHS)[1] for researchers to compare their WHS methods using the same dataset, where most of the top-ranked methods are based on Convolutional Neural Networks (CNNs).

Despite the advances in deep learning techniques and many successful applications for medical images [6], [7], there are still some open challenges to address when adopting 3D CNNs in clinical practice. The first challenge is the limited number of training samples. Due to availability, cost of manual annotation, and privacy issues, it is usually difficult to collect enough training data, especially for 3D images, to cover the variances across subjects and imaging acquisitions. As a result, it is very likely to obtain a small set of labeled examples in

This work was supported in part by the Beijing Nova Program under Grant Z201100006820064, in part by the National Key Research and Development Project of China under Grant 2020YFC2004800, in part by the STCSM under Grant 19511121400, and in part by the Beijing Postdoctoral Research Foundation. Corresponding author: Qing Xia (e-mail: xiaqing@sensetime.com).

W. Wang, Z. Hu, Z. Li, N. Huang are with Sensetime Research, China. Q. Xia is with Sensetime Research, China and the Department of Software, Tsinghua University, China. Z. Yan is with SenseBrain Technology Limited LLC, Princeton, NJ 08540 USA. Y. Wu is with Chinese PLA General Hospital, China. Y. Gao is with the Department of Software, Tsinghua University, China. D. Metaxas is with the Department of Computer Science, Rutgers University, NJ, 08854 USA. S. Zhang is with SenseTime Research, Shanghai 200233, China, and also with the Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai 200240, China.

[1] https://zmiclab.github.io/projects/mmwhs/

many applications, especially at the early stages or for rare diseases [8]. Learning a robust model from a few labeled data is very challenging but valuable for those use cases with difficulties in data and annotation collection, and meaningful in some preliminary studies with limited resources. Another practical challenge is the GPU memory limitation for 3D CNNs. In order to fully utilize 3D information in CT images, 3D CNN is a more straightforward choice than 2D CNNs. However, it is not trivial to directly use 3D CNNs for segmentation because of the huge memory requirements of the intermediate feature maps for a 3D input. Downsampling the input or cropping sub-volumes are two commonly used strategies for training a 3D CNN on large images. The downsampling operations sacrifice accuracy along boundaries, while the cropping strategies have limited observation of global information and thus may lead to inconsistency across sub-volumes.

To overcome the above-mentioned challenges and limitations, we propose a few-shot learning framework by combining ideas of semi-supervised learning and self-training. In our framework, we first adopt a teacher-student model in the initial semi-supervised learning stage and obtain pseudo labels for unlabeled data. In order to select reliable pseudo labels for the next learning stage, a shape-constrained network is built to estimate their qualities. Then, we design a self-training method to update pseudo labels and the segmentation model by using downsampling and cropping strategies alternately, which can make good use of the complementary global and local perspectives to avoid accumulative bias.

We evaluate our method on the CTA (CT angiography) dataset of MM-WHS 2017 Challenge, which includes 20 labeled CT data in the training set and 40 unlabeled samples in the testing set. With the help of more unlabeled data, our method can achieve fully automatic WHS with a Dice Coefficient of 0.917, a Jaccard of 0.848, and a Hausdorff Distance of 15.709 $mm$ on the 40 testing images in the challenge, which outperforms the state-of-the-art methods with distinct improvements. More importantly, our method is able to achieve comparable performance even with very few labeled data and obtain satisfying results on unseen data acquired from different sources. In addition, we apply the proposed method to two 4D (3D+T) CTA data and demonstrate its potential in clinical practice by quantitatively evaluating the cardiac functionality via the time-varying volume of LV and accurate Ejection Fraction.

The main contributions of this paper include: (1) an easy-to-implement few-shot learning method by combining mean teacher model and self-training in a cascaded framework; (2) an effective shape-constrained network to estimate the quality of pseudo-labels; (3) a flexible integration of global and local perspectives for high-resolution 3D data by using downsampling and cropping strategies in an alternative way; (4) quantitative validation on a challenge dataset and qualitative evaluation on a private multi-center dataset.

The rest of the paper is organized as follows. Section II introduces some related work. Section III provides details of the proposed method. Section IV reports the experimental settings and results, followed by discussions. We conclude this work in Section V.

## II. Related work

### A. Semi-supervised image segmentation

In the computer vision domain, there are many semi-supervised methods (e.g., [9], [10]) and self-training approaches (e.g., [11]). For example, Xie et al. [11] proposed a self-training framework with a noisy student to improve the ImageNet classification. Recently, some semi-supervised methods have been applied for medical image analysis. Bai et al. [12] proposed a self-training based method for cardiac MR image segmentation. They updated the network parameters and pseudo labels of unlabeled data in an alternative way. The pseudo labels of unlabeled data are predicted by the network and post-processed by a conditional random field (CRF). Li et al. [13] used a self-training method to refine the target model iteratively by learning from previous predictions of unlabeled data to detect cells in histopathological images. Besides pseudo-label guided methods, Baur et al. [14] utilized manifold embedding to minimize the feature distance between labeled and unlabeled samples. Cui et al. [15] and Li et al. [16] adapted a mean teacher model [17] for their segmentation tasks by using a consistency loss to minimize the difference between a teacher model and a student model for unlabeled data. Some other methods use an adversarial loss to let the segmentation model learn from a discriminator for unlabeled data [18], [19]. No matter how the predictions of unlabeled data are used in different approaches, the noisy labels in the prediction are inevitable and thus affect the training process.

To better utilize the noisy prediction of unlabeled data, some studies have been exploring the certainty or confidence estimation. Zou et al. [20] used a hyper-parameter to control class-balanced pseudo-label selection to determine the most confident samples in each class for robust self-training. Then, they proposed two types of regularization to further improve the confidence guided self-training [21]. Lately, Yu et al. [22] utilized an uncertainty-aware consistency loss to estimate voxel-level uncertainty in the teacher model predictions and let the student model only learn from the confident voxels. Nie et al. [23] designed an attention mechanism by using an adversarial network to learn the confidence map.

Despite the great achievements of semi-supervised methods, only very few data are available in many cases, especially for medical applications. Therefore, few-shot learning (FSL) methods are attracting more attention recently, which can generalize to a task containing only a few labeled samples by using some prior knowledge [24], [25]. There are many different ways of utilizing prior knowledge, for example, by adopting a pre-trained model learned from other larger datasets. In computer vision, most FSL works focus on classification tasks. So far, only a few studies have investigated the FSL for medical image segmentation. Dietlmeiera et al. [26] leveraged convolutional features from a pre-trained VGG-16 network to train a binary gradient boosting classifier from two 1728×2022 images to classify cell pixels in electron microscopy images. However, a pre-trained model is generally unavailable for MRI or CT scans. Mondal et al. [27] proposed a method based on Generative Adversarial Networks (GANs) to segment the brain in 3D multi-modal MRI with

only 1 or 2 labeled samples and some unlabeled images. Zhao et al. [28] presented a learning-based data augmentation method for synthesizing labeled medical images from only a single labeled data and 100 unlabeled samples, and trained a supervised model with these generated examples for brain MRI segmentation. Roy et al. [29] proposed a novel few-shot framework to segment multiple organs in volumetric CT images, which can incorporate strong interactions at multiple locations to ease the training of the segmenter without the need for any pre-trained model.

Inspired by these previous works, we propose our cascaded learning framework by combining the mean teacher model and self-training. Since boundaries and spatial information are vital for multi-label segmentation tasks, we design a shape-constrained network to estimate the pseudo label quality on the subject level instead of the voxel level to help in learning a robust model from only a few labeled data.

### B. CNN-based 3D cardiac segmentation

Among those deep learning-based methods for cardiac segmentation, most are designed for ventricle segmentation, especially in MR and ultrasound domains [30]. In this work, we focus on WHS for 3D CTA images. In order to fully utilize 3D information in 3D image segmentation tasks with limited GPU memory, different approaches have been presented. In some work, downsampling and cropping techniques are used for 3D CNNs, and other work use multi-view 2D CNNs and different fusion strategies to combine complementary information from different views, which is also called 2.5D segmentation.

To name a few approaches using downsampling, Payer et al. [31] proposed a two-step segmentation framework by using a localization CNN in coarse resolution and a segmentation CNN to segment the fine details in the detected small region of interest (ROI). Their method achieved the best performance on the CT segmentation task in the MM-WHS 2017 challenge. Tong et al. [32] used a similar two-step approach and extracted multi-modality features by fusing MRI and CTA images. Isensee et al. [33] also used a similar cascade U-Net in their segmentation framework and achieved promising performances for different segmentation tasks in the Medical Segmentation Decathlon challenge[2]. In summary, downsampling strategies need some kind of refinement as extra steps to further smooth the high-resolution boundaries.

Another representative solution is to train networks by using cropped sub-volumes [34], [35], and then merge the predictions via some fusion strategies. For example, Yang et al. [36], [37] used random cropped patches to train the 3D fully convolutional network (FCN) and adopted sliding window and overlap-tiling stitching strategies to generate predictions for the whole heart volume.

Besides, Wang and Smedby [38] proposed a way to combine three orthogonal 2D U-Nets for 2.5D segmentation and refine the segmentation by a shape context estimation. Similarly, Mortazi et al. [39] used an adaptive fusion strategy to combine the probability maps of multi-object multi-planar CNNs. Zheng et al. [40] proposed a heterogeneous feature aggregation network to exploit complementary information from multiple views of 3D cardiac data by using asymmetrical 3D kernels.

In this work, we utilize both cropping and downsampling strategies in the cascaded learning to explore complementary information in different levels of receptive field and resolution.

### C. Shape prior for segmentation

In many applications, incorporating prior information about anatomies are useful to improve the performance of image segmentation algorithms as summarized in a recent survey [41]. Shape prior, one of the many forms of prior information, provides a powerful semantic description for targeted objects in an image. In the WHS task, the shape of the heart and its substructures can vary from one subject to another or even over time. Many statistical models have been proposed to capture the intra-class variation of shapes, for example, active shape models [42], [43], sparse shape composition [44]–[46], and so on. Some recent methods [47], [48] have tried to incorporate shape priors into segmentation networks in supervised learning to encourage the prediction to be similar to the learned shape and ground truth. Dalca et al. [49] proposed a way of learning anatomical priors from unpaired segmentation images for unsupervised segmentation. In this paper, we designed an auto-encoder network to learn shape priors for the sake of pseudo label selection in a semi-supervised learning framework.

## III. METHODS

The proposed cascaded learning framework for WHS is illustrated in Fig. 1 and summarized in Algorithm 1. It consists of three learning stages ($R = 3$), shown as three gray blocks (from top to bottom) on the left side of Fig. 1. The inputs to the first learning stage include labeled and unlabeled data. Because the number of labeled data is so limited, we use a semi-supervised learning network, specifically a mean teacher model [17], to utilize the unlabeled data in this initial stage. Note that conventional self-training methods generally learn an initial model from labeled data only (e.g., [11]). Since learning a deep network from a small dataset can lead to the over-fitting problem, the initial semi-supervised learning stage should be a better choice, which is validated in our experiments. The learned initial model is then used to generate pseudo labels for the unlabeled data. To estimate the quality of pseudo labels and control how the pseudo labels are used in later self-training stages, we train an auto-encoder network to learn shape priors and to measure the similarity between a prediction and its shape-constrained reconstruction, shown in the right block of Fig. 1. In the second and third self-training stages, the inputs are labeled data, unlabeled data with pseudo labels. We utilize both high-quality and low-quality pseudo labels in self-training, but with different weights to train a new model. Notably, the second and third stages are different, because we employ different downsampling and cropping strategies on the input to avoid an accumulation of segmentation bias. During self-training (two rounds in Fig. 1), the quality of pseudo labels becomes better, and the unlabeled data with better pseudo labels can help to learn a more general segmentation model.

[2]http://medicaldecathlon.com/

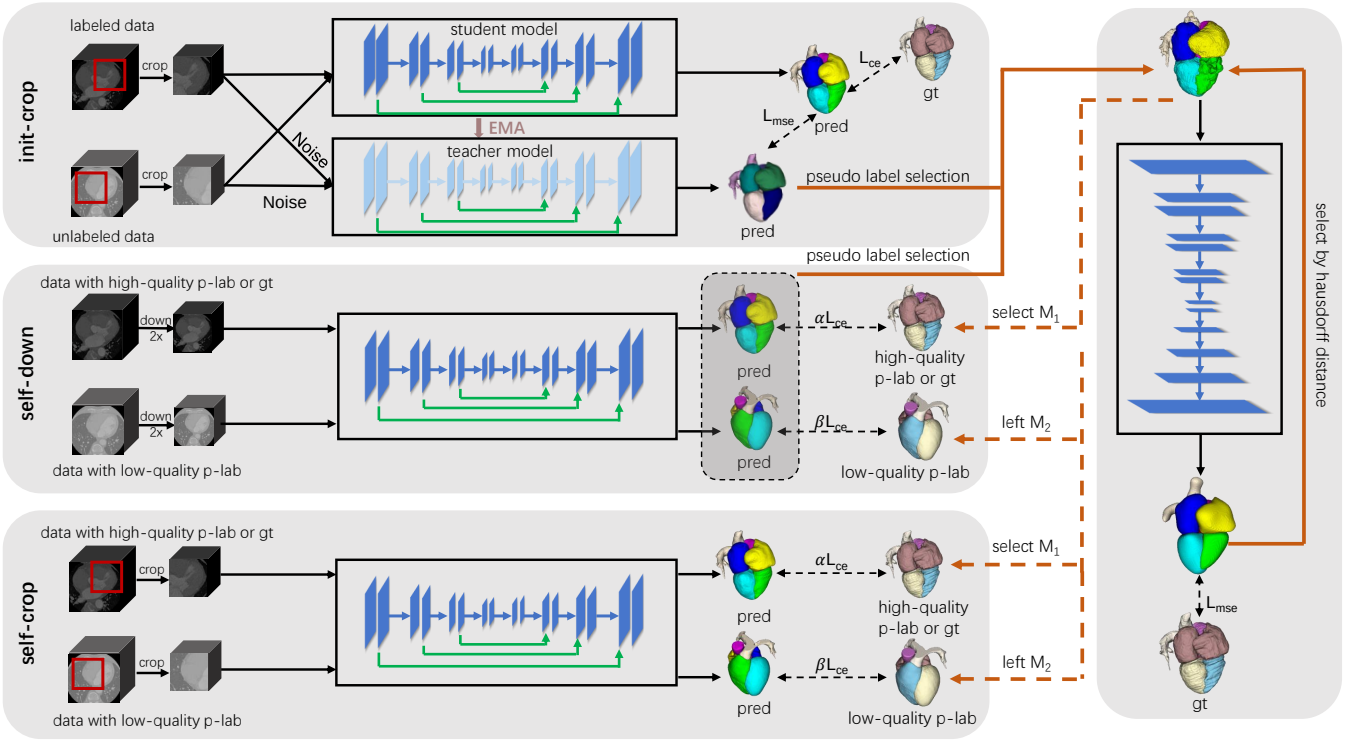



Fig. 1: The proposed few-shot learning framework for WHS. The three-stage learning process is illustrated as three blocks on the left side: initial learning stage (init-crop), and two self-training stages by using downsampled inputs (self-down) and cropped inputs (self-crop), respectively. The shape-constrained network on the right side aims to estimate the quality of pseudo labels of unlabeled data between learning stages. The selected high-quality pseudo labels are treated the same as the ground-truth labels with higher learning weight than the low-quality ones in the next stage. Different color maps are used for 3D heart models of various meanings, i.e., the ground truth of labeled data and the predictions or pseudo labels of the unlabeled data.

The proposed framework provides a semi-supervised solution not only to WHS in CTA. It is robust when only a few labeled data are available. The number of learning stages and labeled data could be different for different use cases. We suggest the maximum number of learning stages, $R$, to be an odd number and at least three. It is because that ending with the local-detail learning stage can provide finer boundaries in prediction. Without loss of generality, the method is presented for WHS by using three learning stages in the few-shot context in this paper. The details of our framework are described in the following subsections.

### A. Initial Semi-supervised Learning

In the beginning, we utilize mean-teacher architecture, a semi-supervised learning method, as the initial stage to start the following self-training. Mean teacher [17] is a self-ensembling method to effectively take advantage of unlabeled data to alleviate the limitation of a small number of labeled data. There are two models in the *mean-teacher* framework, serving roles as student and teacher, respectively. The network parameters of the student model are updated by gradient descent while the parameters of the teacher model are updated as an exponential moving average (EMA) of the student's parameters. The goal of the mean-teacher model is to minimize the following objective function:

$$\min_{\theta} \sum_{i=1}^{N} \mathcal{L}_{ce}(f(x_i;\theta), y_i) + \lambda \sum_{i=1}^{N+M} \mathcal{L}_{mse}(f(x_i;\theta), f(\hat{x_i};\hat{\theta})), \tag{1}$$

where $N$ and $M$ are the number of labeled data and unlabeled data (assuming $N << M$), respectively. $\mathcal{L}_{ce}$ denotes cross-entropy loss, and $\mathcal{L}_{mse}$ represents mean square error loss. The input of the teacher model, $\hat{x_i}$, is a permuted version of $x_i$ for the student model by adding Gaussian noise. $\hat{\theta}$ represents the parameters of the teacher model, which is the EMA of the student model's parameters $\theta$. Besides the parameter values, both teacher model and student model share the same network architecture, and thus the operations are the same. This objective function consists of two main parts. The first term is a segmentation loss from the ground-truth labels, $y_i$, of the labeled data only. The second one is a consistency loss from the targets obtained from the teacher model for all data. After training, we use the teacher model to infer segmentation results as pseudo labels for the unlabeled data and then use them to supervise the subsequent self-training stages after pseudo label selection.

Considering the memory limitation for high-resolution 3D image volume, we need to train the network by using either downsampled input or cropped sub-volumes. Generally

**Algorithm 1:** Cascaded learning of WHS

**Input:** Labeled data $\{(x_1, y_1), ..., (x_N, y_N)\}$, unlabeled data $\{x_{N+1}, ..., x_{N+M}\}$

I. Learn an initial semi-supervised model $\theta_0$ from both labeled and unlabeled data by optimizing Eq. 1 and using randomly cropped sub-volumes of $x_i$ as input;

II. Learn a shape-constrained model $\theta_{sc}$ according to Algorithm 2;

/* Then, the self-training: */

**for** $r = 1 : R$ **do**

  /* $1 \leq r < R$ ($R$: maximum learning stages and an odd number) */

  III. Generate pseudo labels $p_i$ for the unlabeled data by using model $\theta_{r-1}$, and select $M_1$ good-quality ones using model $\theta_{sc}$ based on Eq. 3;

  **if** *$r$ is odd* **then**

    IV. Use downsampled whole images as input to learn a model $\theta_r$ by optimizing Eq. 4;

  **else**

    IV. Use randomly cropped sub-volumes as input to learn a model $\theta_r$ by optimizing Eq. 4;

  **end**

**end**

**return** *The learned model* $\theta_{R-1}$.

**Algorithm 2:** Training of shape-constrained network

**Input:** Labeled data $\{(x_i, y_i) | i \in (1, ..., N)\}$ and the learned teacher model $\theta_0$ in the initial stage

1. Off-line data augmentation: generate $K$ different predictions $\{p_{i,k}\}$ from each $x_i$;

2. Learn the shape-constrained network $\theta_{sc}$ by optimizing Eq. 2 with on-line data augmentation;

**return** *The learned model* $\theta_{sc}$.

speaking, learning from downsampled input may require more annotated data because one volume can only serve as one training sample in this case compared to that in the cropping approaches where one image provides multiple training samples by the cropped sub-volumes. Because a very limited number of labeled samples may not provide enough diversity for supervised segmentation loss, we randomly crop the 3D image into sub-volumes as input in this initial training stage.

### B. Pseudo label selection

In practice, the unlabeled data may be acquired by different machines with different protocols and thus have different appearance patterns. A model learned from a limited number of labeled samples may not produce satisfying predictions for the unlabeled data. To select reliable pseudo labels at the initial semi-supervised learning stage or self-training stage, we need to estimate the quality of predictions.

Uncertainty estimation is a commonly used strategy, which assumes that the certainty of prediction can estimate the prediction accuracy. However, it may not work well for use cases with only a few labeled samples because the data variance is so limited that the trained model may produce some wrong labels with high confidence on unseen data and vice versa. Besides, for multi-label segmentation task, boundaries and spatial information of substructures are useful. Since the challenging or uncertain regions are usually along boundaries, a voxel-level uncertainty estimation may lead to isolated sub-regions and thus it is likely that some important spatial information is missing in the filtered pseudo labels.

In this paper, we propose a new method for subject-level pseudo label selection based on shape prior, which will select unlabeled samples with an overall acceptable prediction for the following learning. We assume that an overall acceptable prediction is more helpful than some certain voxels in such few-shot learning for robust segmentation and the accuracy could be further improved as more and more reliable pseudo labels are included in the self-training.

Our selection method is based on a shape prior estimation. We use a convolutional auto-encoder network to directly learn the shape and position information of the cardiac structures. A diagram of the network is shown in the right box of Fig. 1. The inputs to the network are the predictions of labeled data, $p_i$, and outputs are reconstructed shapes. Since the number of available labeled data $N$ could be very small, we introduce various augmentation strategies to increase the training samples by $K$ times to train the auto-encoder network. This model learns the shape prior by optimizing:

$$\min_{\theta_{sc}} \sum_{i=1}^{N} \sum_{k=1}^{K} \mathcal{L}_{mse}(f_{sc}(p_{i,k}; \theta_{sc}), y_i), \tag{2}$$

where $f_{sc}$ and $\theta_{sc}$ denote the shape-constrained network and its parameters, and $p_{i,k}$ is the augmented prediction. We employ off-line and on-line augmentation strategies. For off-line data augmentation, we collect predictions by using several teacher model parameters saved at different epochs. For each teacher model, we inject random Gaussian noise into a random region in the input image and employ optional connected-component post-processing to obtain various predictions. After collecting $K$ variants $\{p_{i,k}\}$ for each $x_i$, we train the shape-constrained network with additional on-line augmentation of $p_{i,k}$, such as random flip, rotation, and scaling. The training process is summarized in Algorithm 2.

By this means, the shape-constrained model can learn to fix incorrect segmentation results (especially for those outlier results) by taking advantage of cardiac structures' prior shape and position information. Furthermore, we estimate the quality of the pseudo labels by measuring the Hausdorff Distance (HD) and Dice score between the predictions and their reconstructed results. As a result, the smaller the HD, the higher its quality. The selection of a high-quality pseudo label consists of three steps. First, we compute statistics of HD and Dice values for the cardiac structures in each subject. Then, we sort the cases according to their average HD ($HD_{mean}$) and select $M/2$ pseudo labels with the smallest $HD_{mean}$'s. Last, we filter out some bad cases that do not meet the following thresholds from the top-$M/2$ cases:

$$Dice_{mean} \geq 0.8 \cap HD_{max} \leq HD_{mean} + 1.8 * HD_{std}, \tag{3}$$

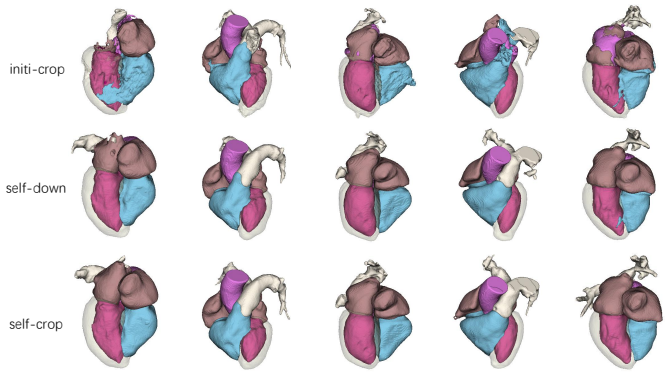



Fig. 2: Evolution of pseudo labels of five unlabeled samples during the cascaded learning. The pseudo labels are becoming better in each column.

where $Dice_{mean}$, $HD_{max}$, and $HD_{std}$ are the average Dice, maximum HD, and standard deviation of HD of cardiac substructures for each subject, respectively. As a result, we have $M_1 (\leq M/2)$ selected pseudo labels. $M_1$ tends to become bigger after more learning stages. The strategy of filtering and selection could be defined differently in other use cases. In this work, because we found that the labels of PA are sparse and inconsistent in different data samples, the statistics are computed excluding PA.

Fig. 2 depicts the evolution of pseudo labels of five examples. It shows that pseudo labels getting improved during the cascaded semi-supervised learning, and the shape-constrained model gradually recognizes them as reliable labels to supervise the model updating in the subsequent learning stage.

### C. Self-training

Our cascaded framework has self-training stages following the initial semi-supervised learning. Because the labeled dataset is too small to be representative in the few-shot context, the trained model could suffer a significant over-fitting problem and a single semi-supervised learning may not be enough to obtain satisfying results.

After pseudo label selection, both high-quality pseudo labels and low-quality pseudo labels are used in the following self-training stage. The objective function of the self-training stage is:

$$\min_{\theta} \alpha \sum_{i=1}^{N+M_1} \mathcal{L}_{ce}(f(x_i;\theta), y_i(p_i)) + \beta \sum_{i=1}^{M_2} \mathcal{L}_{ce}(f(x_i;\theta), p_i), \tag{4}$$

where $M_1$ and $M_2$ are the numbers of unlabeled data with high-quality pseudo labels and low-quality pseudo labels, respectively, such that $M_1 + M_2 = M$. $y_i(p_i)$ equals to $y_i$ for a labeled sample or the pseudo label for an unlabeled sample. Weights $\alpha$ and $\beta$ ($\alpha > \beta$) control the contribution of reliable and unreliable pseudo labels. In this way, more and diverse supervisions are provided for the model learning and the noisy pseudo labels cannot mislead the model much due to the lower weight $\beta$. After training, predictions are renewed for all unlabeled data and pseudo label estimation is conducted again to update the training set.

During the self-training stages, it may be possible to update the pseudo labels for unlabeled data and achieve better and better model after iterations. However, simply repeating could also lead to accumulative bias errors resulting in performance drops. In this work, we use a simple yet effective strategy to update the pseudo labels during self-training. Specifically, after the initial learning stage, we adopt two different ways, namely global-context learning and local-detail learning, to train our 3D CNN alternatively in cascaded stages. Note that during one single learning stage, only one of these two strategies is used. Although the self-training could continue for several rounds, in this paper we only conduct one global-context learning stage and one local-detail learning stage.

*1) Global-context learning stage:* A model is learned to segment cardiac structures from a global perspective by using downsampled images as input. In this way, global context such as shape and relative positions can be better captured to produce more robust segmentation results, in particular preventing from isolated regions for a continuous object and inconsistent label inside the same object.

*2) Local-detail learning stage:* Because of the different characteristics of downsampling and cropping strategies, we utilize them alternatively in the cascaded framework to avoid accumulative bias. Therefore, a global-context learning stage is followed by local-detail learning by using cropped sub-volumes. This stage can help to further refine the local boundary details in the prediction.

## IV. Experiments

### A. Dataset

We collected three CTA datasets. The first one is the CTA dataset from the MM-WHS 2017 Challenge [4]. This dataset consists of 20 labeled data for training and 40 unlabeled data for testing. The slices were acquired in the axial view. The in-plane resolution is about 0.434 × 0.434 $mm^2$ and the average slice thickness is 0.596 $mm$. Each of the training data has seven substructures of the heart labeled, including left ventricle (LV), right ventricle (RV), left atrium (LA), right atrium (RA), myocardium of LV (MYO), ascending aorta (AO), and pulmonary artery (PA). The second dataset contains 128 unlabeled CTA data from six different centers, whose average in-plane resolution is 0.384 × 0.384 $mm^2$ (range from 0.275 × 0.275 $mm^2$ to 0.563 × 0.563 $mm^2$) and the average slice thickness is 0.469 $mm$ (range from 0.250 $mm$ to 0.625 $mm$). We train the segmentation models by using a combination of different numbers of labeled data and unlabeled data. We evaluate the models quantitatively on 40 testing data of MM-WHS 2017 by using the open-source evaluation code provided by the organizer and also qualitatively on another 40 private testing data, which come from the same centers as the 128 unlabeled data in training. The third dataset has two 4D CTA data used for dynamic analysis of heart function.

### B. Experimental settings

Since we focus on few-shot semi-supervised learning in this study, we design several experiments by using labeled samples

TABLE I: Results of few-shot learning (4 labeled and 64 unlabeled data) on the MM-WHS testing dataset by using the models learned in different stages with different learning strategies, including average Dice, Jaccard, and HD ($mm$) for substructures, and average $\pm$ std for 'WHS' scores. The bold fonts indicate the best values of each column in the sub-sections divided by the middle rules.

| | Methods | Substructures | | | | | | | | WHS |
|---|---|---|---|---|---|---|---|---|---|---|
| | | LV | RV | LA | RA | MYO | AO | PA | mean | |
| Dice | baseline-crop | 0.881 | 0.856 | 0.676 | 0.829 | 0.659 | 0.853 | 0.622 | 0.768 | 0.769±0.172 |
| | init-crop | 0.903 | 0.830 | 0.776 | 0.766 | 0.745 | 0.795 | 0.629 | 0.778 | 0.807±0.114 |
| | self-down-r1 | 0.914 | 0.847 | 0.821 | 0.864 | 0.811 | 0.858 | 0.695 | 0.830 | 0.851±0.104 |
| | self-crop-r2 | 0.933 | 0.882 | 0.867 | **0.909** | 0.824 | 0.920 | 0.789 | 0.875 | 0.886±0.029 |
| | self-down-r3 | 0.916 | 0.845 | 0.845 | 0.885 | 0.803 | 0.891 | 0.721 | 0.844 | 0.858±0.059 |
| | self-crop-r4 | **0.935** | **0.890** | **0.872** | 0.895 | **0.831** | **0.943** | **0.798** | **0.881** | **0.890±0.038** |
| | self-fcrop-r1 | 0.928 | **0.883** | 0.840 | 0.866 | **0.820** | 0.886 | 0.739 | 0.852 | 0.869±0.083 |
| | self-fcrop-r2 | 0.934 | **0.883** | 0.858 | **0.893** | 0.812 | 0.922 | 0.749 | **0.864** | 0.878±0.046 |
| | self-fcrop-r3 | 0.935 | 0.882 | **0.860** | 0.885 | 0.818 | 0.926 | 0.729 | 0.862 | **0.879±0.042** |
| | self-fcrop-r4 | **0.937** | 0.851 | 0.858 | 0.873 | 0.803 | **0.937** | **0.754** | 0.859 | 0.869±0.041 |
| | self-down-r1(w/o sel) | **0.915** | 0.824 | **0.800** | **0.854** | 0.771 | 0.828 | 0.688 | 0.811 | 0.830±0.105 |
| | self-crop-r2(w/o sel) | 0.892 | **0.874** | 0.797 | 0.837 | **0.795** | **0.872** | **0.724** | **0.827** | **0.845±0.134** |
| Jaccard | baseline-crop | 0.810 | 0.753 | 0.574 | 0.758 | 0.539 | 0.812 | 0.498 | 0.678 | 0.650±0.191 |
| | init-crop | 0.828 | 0.713 | 0.654 | 0.674 | 0.621 | 0.731 | 0.505 | 0.675 | 0.689±0.133 |
| | self-down-r1 | 0.850 | 0.739 | 0.714 | 0.794 | 0.697 | 0.801 | 0.574 | 0.738 | 0.750±0.119 |
| | self-crop-r2 | 0.877 | 0.790 | 0.768 | **0.838** | 0.708 | 0.857 | 0.667 | 0.787 | 0.796±0.045 |
| | self-down-r3 | 0.850 | 0.736 | 0.739 | 0.812 | 0.679 | 0.824 | 0.597 | 0.748 | 0.756±0.080 |
| | self-crop-r4 | **0.880** | **0.803** | **0.778** | 0.830 | **0.717** | **0.894** | **0.683** | **0.798** | **0.804±0.058** |
| | self-fcrop-r1 | 0.868 | **0.792** | 0.742 | 0.798 | 0.701 | 0.841 | 0.615 | 0.765 | 0.776±0.103 |
| | self-fcrop-r2 | 0.879 | **0.792** | 0.758 | **0.826** | 0.690 | 0.868 | 0.620 | **0.776** | 0.786±0.066 |
| | self-fcrop-r3 | 0.881 | 0.791 | **0.760** | 0.815 | **0.699** | 0.876 | 0.609 | **0.776** | **0.787±0.062** |
| | self-fcrop-r4 | **0.884** | 0.743 | 0.756 | 0.793 | 0.681 | **0.885** | **0.624** | 0.767 | 0.771±0.061 |
| | self-down-r1(w/o sel) | **0.848** | 0.716 | 0.685 | **0.779** | 0.646 | 0.754 | 0.560 | 0.713 | 0.720±0.124 |
| | self-crop-r2(w/o sel) | 0.828 | **0.780** | **0.692** | 0.770 | **0.679** | **0.827** | **0.611** | **0.741** | **0.749±0.147** |
| HD ($mm$) | baseline-crop | 11.069 | 27.104 | 15.726 | 36.433 | 14.340 | 10.361 | 20.898 | 19.419 | 46.229±29.285 |
| | init-crop | 10.814 | 21.036 | 18.644 | 21.060 | 18.071 | 12.681 | 16.735 | 17.006 | 33.537±16.188 |
| | self-down-r1 | 7.354 | 14.478 | 12.136 | 15.792 | 12.756 | 7.414 | 12.900 | 11.833 | 21.742±13.415 |
| | self-crop-r2 | 6.558 | 12.655 | **10.805** | **15.453** | 9.544 | 5.229 | 9.493 | 9.962 | **18.245±7.511** |
| | self-down-r3 | 6.264 | 11.555 | 11.309 | 16.989 | 10.951 | 6.405 | 12.592 | 10.866 | 21.462±13.671 |
| | self-crop-r4 | **6.157** | **11.006** | 11.379 | 15.511 | **8.348** | **4.783** | **9.010** | **9.456** | 18.539±7.625 |
| | self-fcrop-r1 | 7.741 | 12.787 | 13.538 | **16.249** | 9.932 | 7.872 | 11.337 | 11.351 | 20.876±11.919 |
| | self-fcrop-r2 | 7.290 | 11.895 | **12.426** | 18.733 | 9.707 | 6.057 | 11.131 | 11.034 | 23.246±13.698 |
| | self-fcrop-r3 | 7.539 | **11.215** | 13.079 | 17.749 | **9.045** | 5.832 | 11.010 | **10.781** | **21.473±10.242** |
| | self-fcrop-r4 | **5.846** | 11.369 | 15.537 | 17.259 | 11.402 | **4.907** | **10.722** | 11.006 | 21.611±8.636 |
| | self-down-r1(w/o sel) | **8.446** | **14.023** | **12.912** | 18.844 | 13.702 | 9.897 | **12.680** | 12.929 | 24.880±16.692 |
| | self-crop-r2(w/o sel) | 9.836 | 16.193 | 14.247 | **17.115** | **11.279** | **7.616** | 13.699 | **12.855** | **24.571±14.595** |

from the MM-WHS dataset and unlabeled samples from our private dataset. Firstly, we randomly select 4 labeled samples and 64 unlabeled CTA scans for training to demonstrate the ability of our method in a few-shot learning context. In ablation studies, we investigate the effect of different numbers of labeled and unlabeled data, as well as the suggested components in our framework. Then, we compare the performance of our best model with other state-of-the-art methods in the MM-WHS challenge. For quantitative evaluation, we report the average Dice score, Jaccard index, Hausdorff Distance (HD) of every substructure, the mean value of scores for the seven substructures (column 'mean' in the tables), and the whole heart segmentation scores as utilized in the MM-WHS challenge (column 'WHS'). Note that, the column 'mean' is different from the column 'WHS'. The WHS scores for Dice and Jaccard are the normalized metrics with respect to the size of substructures while WHS HD is the maximum HD value of all substructures [50]. We report standard deviations for column WHS only in the tables due to the limited space. At last, we demonstrate the performance of our best model in a practical study of heart function using 4D CT data.

### C. Implementation details

A tailored V-Net [34], [51] is employed as the backbone for our framework. The encoder part of the network consists of five scales connected by max-pooling layers. The first scale has one convolutional layer converting the input into 16 channels. There are two or three convolutional layers in the next four scales, and the number of feature channels doubles at each max pooling. We use IBN [52] in the first three scales to increase both modeling and generalization capacity. Similar to the encoder, the decoder part has four upsampling layers and finally outputs an 8-channel prediction.

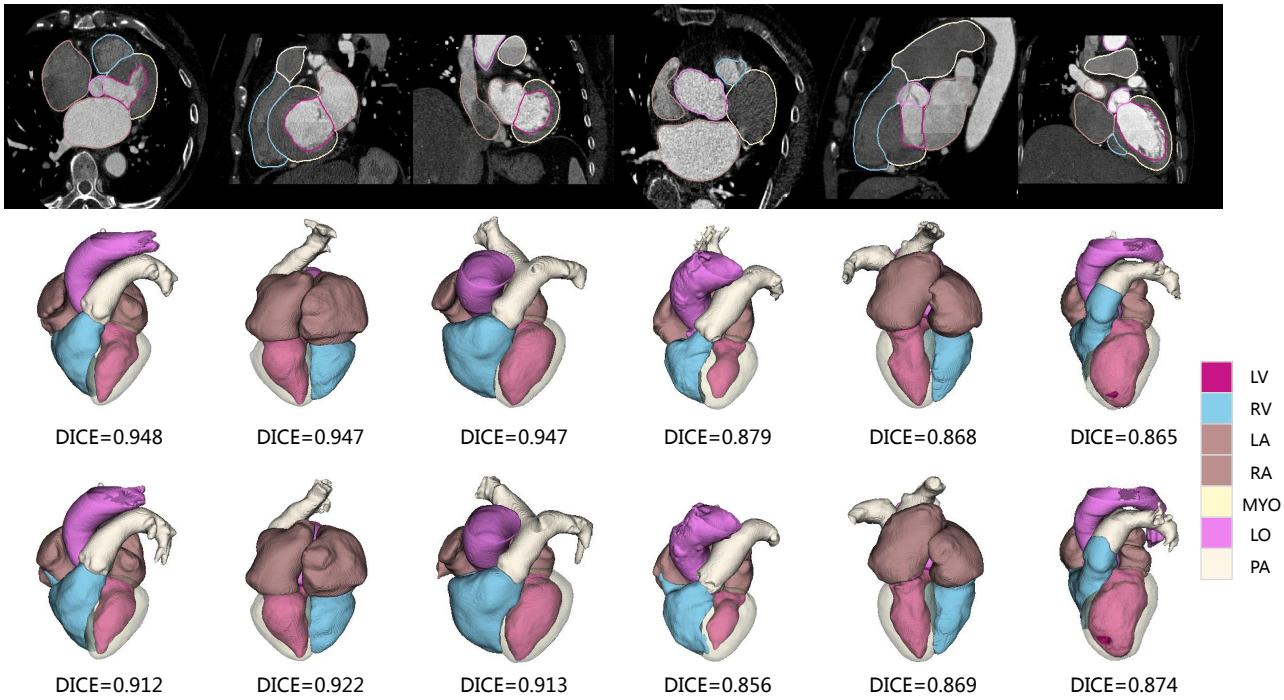



Fig. 3: Qualitative and quantitative results (the WHS Dice) of six MM-WHS testing data (in 6 columns) achieved by the proposed method using 4/16 labeled and 64 unlabeled data. The first row: one 2D slice of each original data from different views overlaid with the segmentation contours. The second row: segmentation results generated by the model trained with 16 labeled data. The third row: segmentation results generated by the model trained with 4 labeled data.

The shape-constrained network consists of one input layer, four encoder blocks, four decoder blocks, and one output layer. The input layer transforms the one-hot prediction into 16-channel feature maps by 3D convolutions (conv3d). Each encoder block has two conv3d layers. The first one has a stride of two and increases the channel number by 16. The second one keeps the same channel number. Each decoder block has one upsampling layer, which doubles the spatial dimensions of feature maps by trilinear interpolation, and one conv3d layer, which decrease the channel number by 16. The output convolution layer finally maps the channel number to 8 in this use case. Every conv3d layer uses a kernel size of 3 and padding of 1, followed by batch normalization and ReLU activation. To alleviate the limitations of GPU memory and keep global shape information, the input to the network is the downsampled prediction.

In data preprocessing, the original cardiac CTA images were resampled to 0.625 $mm$ on three axes. The intensities were normalized by 2048 and clamped between -1 and 1. We used a pre-trained localization network to find the center of the heart and then cropped an ROI of 288×288×288 from each preprocessed image for the experiments. In all learning stages, including the training of shape-constrained network, the size of the input 3D volume was 144×144×144. Several data augmentation techniques, such as scaling, rotation, flipping, and elastic deformation, were randomly performed to increase data variance. In order to control the balance between the supervised loss and unsupervised consistency loss in Eq. 1, we followed [17] to use a ramp-up coefficient $\lambda(T) = 10 * e^{-5(1-T)^2}$, where $T$ changes linearly from zero to one during the ramp-up period. We set $\alpha = 0.8$ and $\beta = 0.2$ for Eq. 4 in our experiments. 4 out of 20 challenge training data were reserved as the validation set to select model while the remaining 16 labeled data were used in training. All networks were trained for 3000 epochs with a minibatch size of 4. All the implementations were performed in Pytorch by using 4 NVIDIA GeForce GTX 1080 Ti GPUs, and it took about 14, 22, and 25 hours, respectively, to train the models in the three learning stages.

### D. Results of Few-shot learning

In many practical applications, a good number of labeled data is not always available. Considering the expenses of annotation, it may be affordable to obtain only a few labeled data. Therefore, the ability to learn from only a few labeled samples has significant practical value. In this section, we demonstrate the performance of our framework in a few-shot learning context. Specifically, we randomly select 4 labeled samples from the MM-WHS dataset and train the cascaded networks with the help of 64 multi-center unlabeled data.

The results on the 40 MM-WHS testing data are shown in Table I. For convenience, we name the results at the initial semi-supervised learning stage init-crop since cropped images are used as input. The global-context learning stage and local-detail learning stage are denoted by self-down and self-crop, respectively, followed by the corresponding stage index. The baseline-crop is a baseline model trained with only 4 labeled data by supervised learning. By comparing the results of

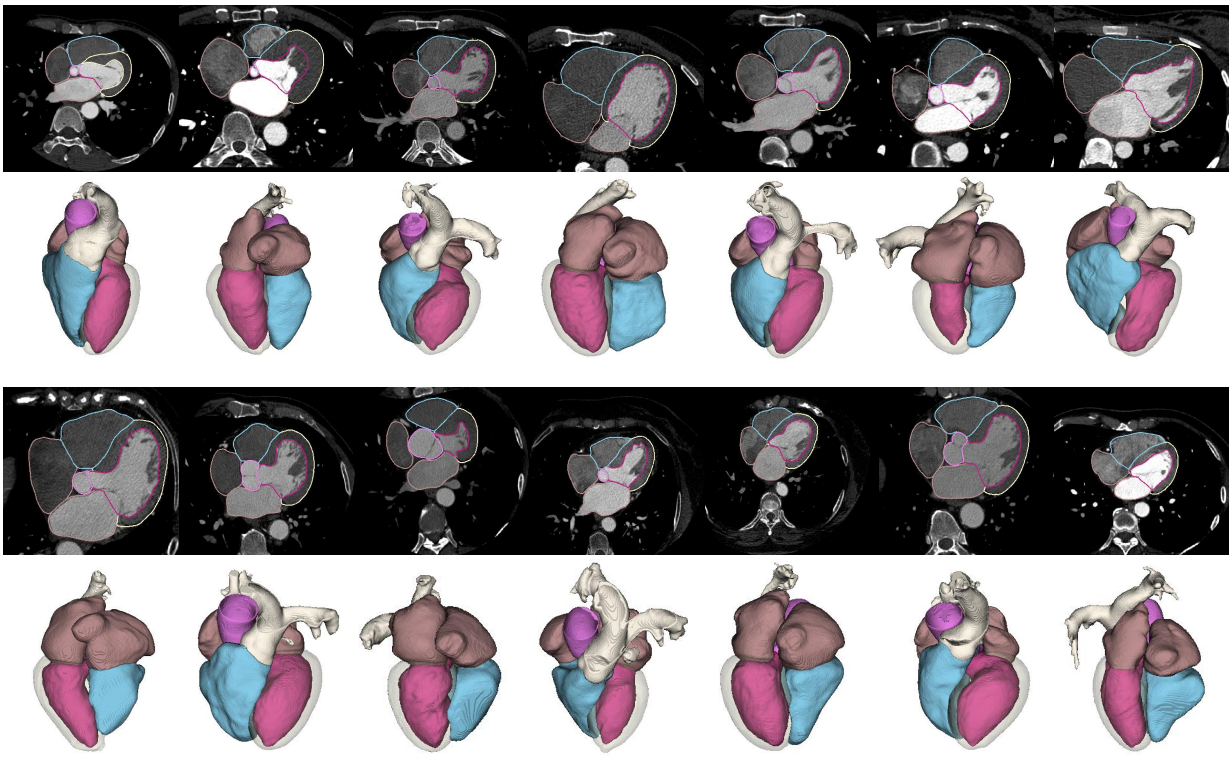


Fig. 4: Qualitative results of 14 samples in our private dataset, generated by the model trained with 4 labeled data and 64 unlabeled data. The 2D views have the segmentation contours overlaid on the original images.

baseline-crop and init-crop, we can see that more unlabeled data in semi-supervised learning can help to achieve a more accurate and robust model. In our self-training stages, the models are learned by using downsampled inputs and cropped inputs alternatively. After the init-crop stage, we can perform four rounds of proposed self-training, which are denoted by self-down-r1, self-crop-r2, self-down-r3, self-crop-r4. We can see from the results in Table I that the overall performance tends to be improved with more learning rounds, which shows the advantage of the proposed cascaded framework. We also observe that the improvement speed decreases as more training rounds. Specifically, by a paired t-test, the differences between the results of self-crop-r2 and self-crop-r4 are not significant in WHS Dice ($p = 0.0808$) and WHS HD ($p = 0.7496$). We conclude that two rounds of self-training after the init-crop should be enough and thus report the results using the models of stage $r = 2$ by default in the following experiments. Since 16 labeled data are available for training, we trained three more self-crop-r2 models by selecting different sets of 4 labeled data in the training. The performances of the trained models are slightly different. The average WHS Dice and WHS HD of all the four models are 0.862 and 21.456 $mm$, respectively, comparable to the results of self-crop-r2 in Table I (details are reported in Table S1 of the supplementary materials).

Fig. 3 shows some qualitative results of six challenge testing samples in 3D views and the corresponding WHS Dice scores are also displayed. The first row shows one 2D slice from each original volumetric image in different views. The second row shows 3D views of corresponding predictions by using a model trained with 16 labeled samples and 64 unlabeled data, which will be detailed later. The third row shows the results by using the model trained with only 4 labeled samples and the same unlabeled data. Each column is for one subject. As we can see that the model trained with only 4 labeled data can still produce reasonable segmentation results, especially for the hard testing cases by the model trained with 16 labeled data.

Moreover, Fig. 4 shows 14 examples of segmentation results in our private dataset. We can see that the proposed method is robust to various image appearances and heart shapes despite only 4 labeled samples are used in training. More segmentation results on our private dataset can be found in Fig. S1 of the supplementary materials.

We conduct several ablation studies to analyze the effects of different components in our framework.

*1) Effect of pseudo-label selection:* After the 'init-crop' model, we train stage-2 and stage-3 models according to Algorithm 1 but without the pseudo-label selection. The results are shown as self-down-r1(w/o sel) and self-crop-r2(w/o sel) in Table I. By comparing with the corresponding self-down-r1 and self-crop-r2 models, which employ the selection strategy, we can see that the pseudo-label selection component benefits the self-training with clearly improved results.

*2) Effect of self-training:* In Table I, we conduct another type of cascaded learning to validate the effect of the proposed self-training method which uses cropped and downsampled input alternatively. This variant of self-training is a series of learning stages that keep using cropped input only. They are denoted by self-fcrop-r1, self-fcrop-r2, self-fcrop-r3, self-fcrop-r4. We have two observations based on the results in Table I. First,

TABLE II: Results on the MM-WHS testing dataset by using the models learned from different numbers of labeled data and 64 unlabeled data, including average Dice, Jaccard, and HD ($mm$) for substructures, and average $\pm$ std for 'WHS' scores.

| | Labeled | Unlabeled | Substructures | | | | | | | | WHS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | LV | RV | LA | RA | MYO | AO | PA | mean | |
| Dice | 4 | 64 | 0.933 | 0.882 | 0.867 | 0.909 | 0.824 | 0.920 | 0.789 | 0.875 | 0.886±0.029 |
| | 8 | 64 | 0.938 | 0.897 | 0.896 | 0.906 | 0.856 | 0.886 | 0.781 | 0.880 | 0.898±0.039 |
| | 16 | 64 | **0.946** | **0.913** | **0.910** | **0.926** | **0.885** | **0.937** | **0.832** | **0.907** | **0.917±0.022** |
| Jaccard | 4 | 64 | 0.877 | 0.790 | 0.768 | 0.838 | 0.708 | 0.857 | 0.667 | 0.787 | 0.796±0.045 |
| | 8 | 64 | 0.885 | 0.815 | 0.815 | 0.835 | 0.756 | 0.843 | 0.677 | 0.804 | 0.817±0.061 |
| | 16 | 64 | **0.898** | **0.841** | **0.837** | **0.864** | **0.799** | **0.887** | **0.732** | **0.837** | **0.848±0.037** |
| HD | 4 | 64 | 6.558 | 12.655 | 10.805 | 15.453 | 9.544 | 5.229 | 9.493 | 9.962 | 18.245±7.511 |
| | 8 | 64 | 7.463 | 10.550 | 10.491 | 14.969 | 8.681 | 7.819 | 9.802 | 9.968 | 20.386±12.980 |
| | 16 | 64 | **6.055** | **9.468** | **8.743** | **13.315** | **7.610** | **4.775** | **7.983** | **8.278** | **15.709±6.684** |

TABLE III: Results of different methods, including supervised (s) and semi-supervised (ss) approaches, on the MM-WHS testing dataset, including average Dice, Jaccard, and HD ($mm$) for substructures, and average $\pm$ std for 'WHS' scores. $^+$ indicates the results are cited from the corresponding paper; $^*$ indicates the results are implemented by us

| | Methods | Substructures | | | | | | | | WHS |
|---|---|---|---|---|---|---|---|---|---|---|
| | | LV | RV | LA | RA | MYO | AO | PA | mean | |
| Dice | Yang [36] (s)$^+$ | - | - | - | - | - | - | - | - | 0.890±0.049 |
| | Wang [38] (s)$^+$ | - | - | - | - | - | - | - | - | 0.894±0.030 |
| | Payer [31] (s)$^+$ | 0.918 | 0.909 | **0.929** | 0.888 | 0.881 | 0.933 | **0.840** | 0.900 | 0.908±0.086 |
| | Li [13] (ss)$^*$ | 0.914 | 0.901 | 0.869 | 0.878 | 0.849 | 0.933 | 0.761 | 0.872 | 0.883±0.110 |
| | Nie [23] (ss)$^*$ | 0.931 | 0.889 | 0.884 | 0.894 | 0.840 | 0.909 | 0.794 | 0.878 | 0.890±0.052 |
| | Xie [11] (ss)$^*$ | 0.939 | 0.881 | 0.896 | 0.877 | 0.862 | 0.921 | 0.795 | 0.883 | 0.895±0.023 |
| | Yu [22] (ss)$^*$ | 0.942 | 0.886 | 0.892 | 0.913 | 0.873 | **0.945** | 0.821 | 0.896 | 0.904±0.029 |
| | ours (ss) | **0.946** | **0.913** | 0.910 | **0.926** | **0.885** | 0.937 | 0.832 | **0.907** | **0.917±0.022** |
| Jaccard | Yang [36] (s)$^+$ | - | - | - | - | - | - | - | - | 0.805±0.074 |
| | Wang [38] (s)$^+$ | - | - | - | - | - | - | - | - | 0.810±0.048 |
| | Payer [31] (s)$^+$ | - | - | - | - | - | - | - | - | 0.832±0.037 |
| | Li [13] (ss)$^*$ | 0.863 | 0.823 | 0.789 | 0.807 | 0.758 | 0.880 | 0.663 | 0.798 | 0.803±0.129 |
| | Nie [23] (ss)$^*$ | 0.874 | 0.803 | 0.799 | 0.812 | 0.744 | 0.852 | 0.686 | 0.796 | 0.806±0.073 |
| | Xie [11] (ss)$^*$ | 0.886 | 0.788 | 0.814 | 0.801 | 0.763 | 0.855 | 0.678 | 0.798 | 0.810±0.037 |
| | Yu [22] (ss)$^*$ | 0.893 | 0.798 | 0.809 | 0.844 | 0.781 | **0.896** | 0.714 | 0.819 | 0.826±0.048 |
| | ours (ss) | **0.898** | **0.841** | **0.837** | **0.864** | **0.799** | 0.887 | **0.732** | **0.837** | **0.848±0.037** |
| HD ($mm$) | Yang [36] (s)$^+$ | - | - | - | - | - | - | - | - | 29.006±15.804 |
| | Wang [38] (s)$^+$ | - | - | - | - | - | - | - | - | 31.146±13.203 |
| | Payer [31] (s)$^+$ | - | - | - | - | - | - | - | - | 25.242±10.813 |
| | Li [13] (ss)$^*$ | 8.749 | 11.868 | 11.610 | 15.927 | 10.262 | 6.159 | 11.335 | 10.844 | 20.660±16.361 |
| | Nie [23] (ss)$^*$ | 7.515 | 12.122 | 15.274 | 17.310 | 8.871 | 6.384 | 9.994 | 11.067 | 23.153±12.624 |
| | Xie [11] (ss)$^*$ | 6.453 | 11.026 | 13.911 | 14.199 | 8.469 | 4.907 | 9.014 | 9.711 | 18.480±6.799 |
| | Yu [22] (ss)$^*$ | 6.589 | 11.308 | 10.123 | 13.385 | 8.557 | 5.266 | 8.163 | 9.056 | 17.427±9.505 |
| | ours (ss) | **6.055** | **9.468** | **8.743** | **13.315** | **7.610** | **4.775** | **7.983** | **8.278** | **15.709±6.684** |

although the variant can also lead to better performance with more stages, the improvement decreases quickly and the performance even drops after three rounds due to the accumulative bias. Second, the proposed method by alternating downsampling and cropping can achieve more robust results than simply repeating cropping. Although downsampling generally has inferior results to the cropping at the same stage (e.g., self-down-r1 vs self-fcrop-r1 and self-down-r3 vs self-fcrop-r3) due to inaccurate boundaries in low-resolution data, the following stage can be improved more significantly in the proposed framework thanks to the complementary information learned in the global context.

*3) Effect of training set size:* To validate the performance of the proposed method for different sizes of the training set, we train models with different numbers of labeled data when the number of unlabeled samples is fixed. The results are reported in Table II (more detailed results for different stages can be found in Table S2 of the supplementary materials).

From the results in Table II, we can observe that the learning process benefits from an increasing number of labeled data in general. It is interesting to see in Fig. 3 and Table II that the results by using 4 labeled data are comparable to those by using 16 labeled data.

On the other hand, an increasing number of unlabeled data does not benefit the learning process with 4 labeled samples (see Table S3 of the supplementary materials). A possible reason is that the unlabeled private dataset has very different appearances from the challenge dataset, and more unlabeled data with higher variance tends to make the learning harder.

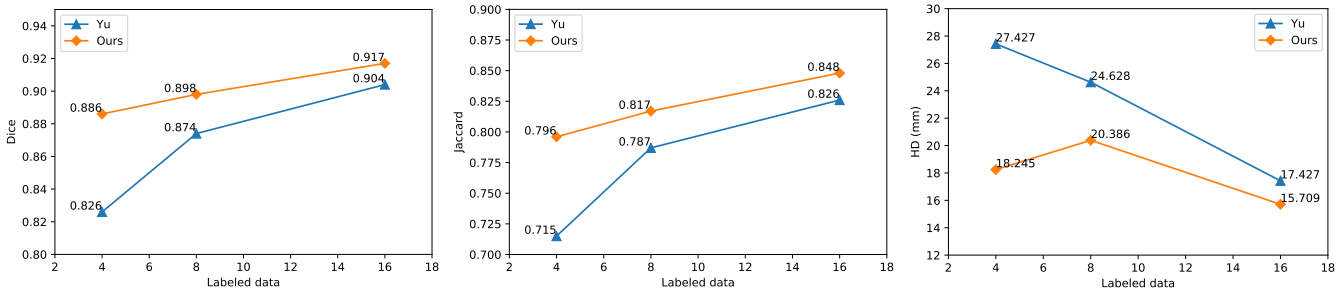



Fig. 5: Comparison of average WHS Dice, WHS Jaccard, and WHS HD achieved by Yu [22] (implemented by us) and the proposed method using different numbers of labeled training data and 64 unlabeled data.

### *E. WHS challenge*

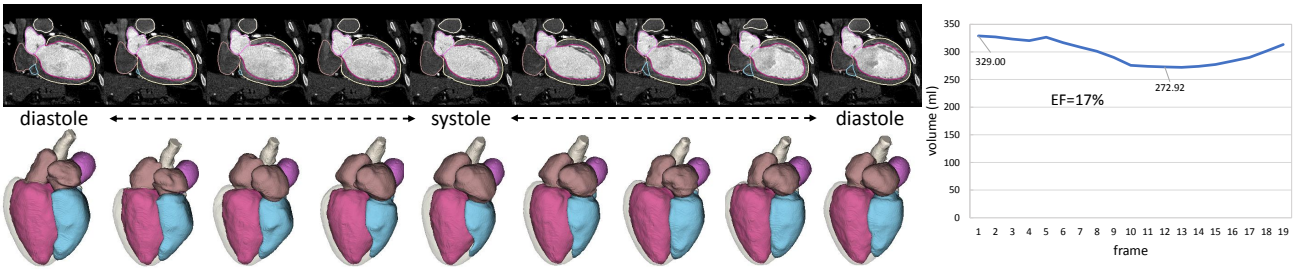
Here, we compare the proposed method with state-of-the-art supervised methods in the MM-WHS 2017 challenge and other recent approaches based on semi-supervised learning. To compare with other methods more fairly, we use the proposed model trained with 16 labeled data and 64 unlabeled data in this subsection. As a supervised learning method, the approach proposed by Payer et al. [31] was the first-ranked submission in the challenge, which uses a coarse-to-fine learning technique. The second-ranked and third-ranked methods [36], [38] are also included in Table III. The proposed method achieves improvements in all metrics and obtains the average WHS Dice of 0.917, WHS Jaccard of 0.848, and WHS HD of 15.709 $mm$. Although the improvements in WHS Dice and WHS Jaccard are slight, the WHS HD is significantly better than [31] (about $38\%$ improvement and $p < 0.0001$ by an unpaired t-test since we don't have their individual results), which indicates that our method is robust enough to handle those hard cases for the other algorithms.

In Table III, we also report the results by some other semi-supervised methods proposed recently (implemented for the WHS task), including two self-training methods [11], [13], a mean teacher-based method [22] and a GAN-based semi-supervised method [23]. In the implementation of the above methods, we used the same 16 labeled and 64 unlabeled data for training and the same 3D CNN-based backbone as in our framework for a fair comparison. Li et al. [13] proposed the self-training method combined with a cooperative-training strategy, i.e., two models are trained from each other's predictions iteratively. Although it prevents the model from getting stuck in its local minimum, the lack of a mechanism for selecting pseudo labels for the unlabeled data affects the experimental results. Nie et al. [23] designed an adversarial network to learn the confidence map to incorporate a voxel-level pseudo-label selection in the learning process. However, due to the limited labeled data in the WHS task, the adversarial network may not be trained sufficiently so that too few confident voxels in unlabeled data are selected to help the learning. The self-training framework proposed by Xie et al. [11] can improve generalization in ImageNet classification by injecting noise such as data augmentation via RandAugment, dropout, and stochastic depth to the student model during training. To extend this 2D classification method for the WHS task, we adopted dropout and stochastic depth for the 3D CNN model and replaced the RandAugment with random augmentation operations such as rotation, flipping, scaling, and elastic deformation. The method of Yu et al. [22] let the student model gradually learn from meaningful and reliable targets by exploiting the voxel-level uncertainty information. Among these comparing methods, Yu [22] achieves the best performance in Table III. By a paired t-test, the results obtained by our method are significantly better than [22] in terms of WHS Dice ($p < 0.0001$) and WHS Jaccard ($p < 0.0001$), which indicate the effectiveness of our method.

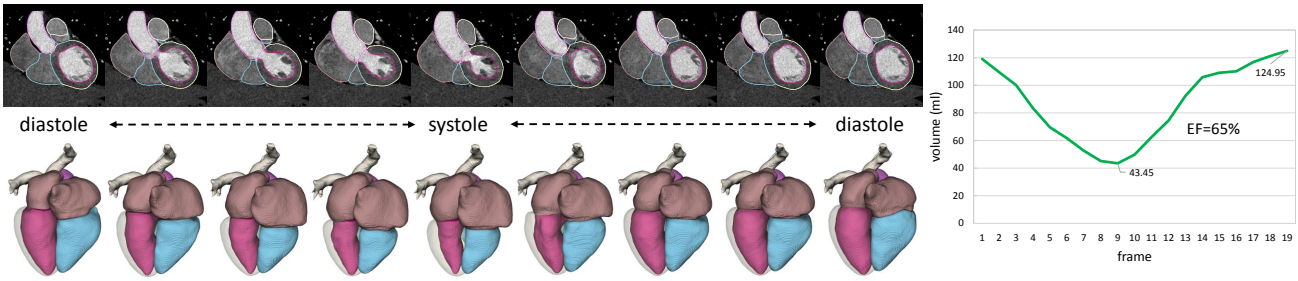
Furthermore, we choose the method by Yu [22] as a representative semi-supervised method and carry out experiments in the few-shot learning context with 4 labeled data and 8 labeled data. We select Yu's method to compare in this setting because it has relatively better results than the other comparing approaches in Table III, and it also achieves state-of-the-art performance on a 3D LA segmentation task [22] by comparing with different semi-supervised methods such as [12], [16], [18], [23]. The comparison of average WHS Dice, WHS Jaccard, and WHS HD are plotted in Fig. 5. We can see that our method shows more advantages with fewer labeled data, which means that it has great potential when only a few labeled data are available.

### *F. 4D case study*

In this section, we apply our model to 4D (3D+T) CTA data for dynamic analysis of heart function. High-resolution 4D CT provides the clinician with high-quality anatomical images. An accurate 4D heart model can further provide intuitive visualization of cardiac motion and quantitative strain analysis to help in clinical practice, such as disease diagnosis and surgery planning. To this end, we employ the proposed method to segment cardiac structures at every frame in a cardiac cycle. Here, we use the model trained with 16 labeled data to do the prediction. The segmentation results of two subjects at 9 out of 19 frames of a whole cardiac cycle are shown in Fig. 6. Fig. 6a is a diseased heart while Fig. 6b shows a normal heart. We can see that our segmentation results are continuous and accurate in both diseased and normal hearts. Based on the 4D segmentation results, the left ventricular volume curves over time are computed and plotted on the right. The diseased LV volume is much larger than the normal one, which could be a sign of chronic hypertension, myocardial infarction, or heart valve disease. Compared with the Ejection Fraction (EF) of




(a) A diseased heart.




(b) A normal heart.

Fig. 6: 4D case study: segmentation results of 9 frames in a cardiac cycle and the evolution curve of the LV volume for a diseased heart (a) and a normal heart (b).

65% of the normal heart, the EF of 17% of the diseased heart reflects some abnormalities of the LV in this subject. The predicted EF values are associate with the values computed from the echocardiogram: 68% for the normal case and 21% for the diseased case. A video representing the 4D whole heart segmentation results can be found in the supplementary materials.

## V. Conclusion

We proposed a novel and easy-to-implement cascaded framework for semi-supervised segmentation and successfully applied in WHS. It combines the ideas of mean teacher and self-training. With the proposed framework, a robust segmentation model can be trained from only a few labeled data. Moreover, an effective shape-constrained network was proposed to select reliable pseudo labels for self-training. We extensively evaluated our method on 40 MM-WHS testing data and some private data. The accurate results of multi-center data demonstrated its effectiveness and generalization ability. Despite the focus on WHS in this paper, our framework can also be extended for wider use. For example, in the near future, we will extend our cascaded framework to 4D cardiac segmentation, where annotating every frame in the whole cardiac cycle is impractical, by exploiting the extra temporal information between consecutive frames, instead of segmenting each frame individually as done in our 4D case study.

## References

[1] S. S. Virani, A. Alonso, E. J. Benjamin, M. S. Bittencourt, C. W. Callaway, A. P. Carson, A. M. Chamberlain, A. R. Chang, S. Cheng, F. N. Delling *et al.*, "Heart disease and stroke statistics—2020 update: a report from the american heart association," *Circulation*, pp. E139–E596, 2020.

[2] D. Kang, J. Woo, C. J. Kuo, P. J. Slomka, D. Dey, and G. Germano, "Heart chambers and whole heart segmentation techniques," *Journal of Electronic Imaging*, vol. 21, no. 1, p. 010901, 2012.

[3] W. Roberts, J. Bax, and L. Davies, "Cardiac CT and CT coronary angiography: technology and application," *Heart*, vol. 94, no. 6, pp. 781–792, 2008.

[4] X. Zhuang, L. Li, C. Payer, D. Štern, M. Urschler, M. P. Heinrich, J. Oster, C. Wang, Ö. Smedby, C. Bian *et al.*, "Evaluation of algorithms for multi-modality whole heart segmentation: An open-access grand challenge," *Medical image analysis*, vol. 58, p. 101537, 2019.

[5] 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.

[6] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[7] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, 2017.

[8] X. Li, L. Yu, Y. Jin, C.-W. Fu, L. Xing, and P.-A. Heng, "Difficulty-aware meta-learning for rare disease diagnosis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 357–366.

[9] A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and I. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," in *Advances in Neural Information Processing Systems*, 2018, pp. 3235–3246.

[10] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 674–12 684.

[11] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 687–10 698.

[12] W. Bai, O. Oktay, M. Sinclair, H. Suzuki, M. Rajchl, G. Tarroni, B. Glocker, A. King, P. M. Matthews, and D. Rueckert, "Semi-supervised learning for network-based cardiac mr image segmentation," in *MICCAI*. Springer, 2017, pp. 253–260.

[13] J. Li, S. Yang, X. Huang, Q. Da, X. Yang, Z. Hu, Q. Duan, C. Wang, and H. Li, "Signet ring cell detection with a semi-supervised learning framework," in *International Conference on Information Processing in Medical Imaging*. Springer, 2019, pp. 842–854.

[14] C. Baur, S. Albarqouni, and N. Navab, "Semi-supervised deep learning for fully convolutional networks," in *MICCAI*. Springer, 2017, pp. 311–319.

[15] W. Cui, Y. Liu, Y. Li, M. Guo, Y. Li, X. Li, T. Wang, X. Zeng, and C. Ye, "Semi-supervised brain lesion segmentation with an adapted mean teacher model," in *International Conference on Information Processing in Medical Imaging*. Springer, 2019, pp. 554–565.

[16] X. Li, L. Yu, H. Chen, C.-W. Fu, L. Xing, and P.-A. Heng, "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[17] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems*, 2017, pp. 1195–1204.

[18] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen, "Deep adversarial networks for biomedical image segmentation utilizing unannotated images," in *MICCAI*. Springer, 2017, pp. 408–416.

[19] A. Chartsias, T. Joyce, G. Papanastasiou, S. Semple, M. Williams, D. Newby, R. Dharmakumar, and S. A. Tsaftaris, "Factorised spatial representation learning: application in semi-supervised myocardial segmentation," in *MICCAI*. Springer, 2018, pp. 490–498.

[20] Y. Zou, Z. Yu, B. Vijaya Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 289–305.

[21] Y. Zou, Z. Yu, X. Liu, B. Kumar, and J. Wang, "Confidence regularized self-training," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5982–5991.

[22] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 605–613.

[23] D. Nie, Y. Gao, L. Wang, and D. Shen, "Asdnet: Attention based semi-supervised deep networks for medical image segmentation," in *MICCAI*. Springer, 2018, pp. 370–378.

[24] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–34, 2020.

[25] J. Lu, P. Gong, J. Ye, and C. Zhang, "Learning from very few samples: A survey," *arXiv preprint arXiv:2009.02653*, 2020.

[26] J. Dietlmeier, K. McGuinness, S. Rugonyi, T. Wilson, A. Nuttall, and N. E. O'Connor, "Few-shot hypercolumn-based mitochondria segmentation in cardiac and outer hair cells in focused ion beam-scanning electron microscopy (FIB-SEM) data," *Pattern Recognition Letters*, vol. 128, pp. 521–528, 2019.

[27] A. K. Mondal, J. Dolz, and C. Desrosiers, "Few-shot 3D multi-modal medical image segmentation using generative adversarial learning," *arXiv preprint arXiv:1810.12241*, 2018.

[28] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca, "Data augmentation using learned transformations for one-shot medical image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 8543–8553.

[29] A. G. Roy, S. Siddiqui, S. Pölsterl, N. Navab, and C. Wachinger, "'squeeze & excite'guided few-shot segmentation of volumetric images," *Medical image analysis*, vol. 59, p. 101587, 2020.

[30] C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, and D. Rueckert, "Deep learning for cardiac image segmentation: A review," *Frontiers in Cardiovascular Medicine*, vol. 7, p. 25, 2020.

[31] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Multi-label whole heart segmentation using cnns and anatomical label configurations," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2017, pp. 190–198.

[32] Q. Tong, M. Ning, W. Si, X. Liao, and J. Qin, "3D deeply-supervised u-net based whole heart segmentation," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2017, pp. 224–232.

[33] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert *et al.*, "nnu-net: Self-adapting framework for u-net-based medical image segmentation," *arXiv preprint arXiv:1809.10486*, 2018.

[34] Q. Xia, Y. Yao, Z. Hu, and A. Hao, "Automatic 3D atrial segmentation from GE-MRIs using volumetric fully convolutional networks," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2018, pp. 211–220.

[35] X. Yang, N. Wang, Y. Wang, X. Wang, R. Nezafat, D. Ni, and P.-A. Heng, "Combating uncertainty with novel losses for automatic left atrium segmentation," *arXiv preprint arXiv:1812.05807*, 2018.

[36] X. Yang, C. Bian, L. Yu, D. Ni, and P.-A. Heng, "3D convolutional networks for fully automatic fine-grained whole heart partition," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2017, pp. 181–189.

[37] ——, "Hybrid loss guided convolutional networks for whole heart parsing," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2017, pp. 215–223.

[38] C. Wang and Ö. Smedby, "Automatic whole heart segmentation using deep learning and shape context," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2017, pp. 242–249.

[39] A. Mortazi, J. Burt, and U. Bagci, "Multi-planar deep segmentation networks for cardiac substructures from MRI and CT," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2017, pp. 199–206.

[40] H. Zheng, L. Yang, J. Han, Y. Zhang, P. Liang, Z. Zhao, C. Wang, and D. Z. Chen, "HFA-Net: 3D cardiovascular image segmentation with asymmetrical pooling and content-aware fusion," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 759–767.

[41] M. S. Nosrati and G. Hamarneh, "Incorporating prior knowledge in medical image segmentation: a survey," *arXiv preprint arXiv:1607.01092*, 2016.

[42] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[43] C. Davatzikos, X. Tao, and D. Shen, "Hierarchical active shape models, using the wavelet transform," *IEEE transactions on medical imaging*, vol. 22, no. 3, pp. 414–423, 2003.

[44] S. Zhang, Y. Zhan, M. Dewan, J. Huang, D. N. Metaxas, and X. S. Zhou, "Towards robust and effective shape modeling: Sparse shape composition," *Medical image analysis*, vol. 16, no. 1, pp. 265–277, 2012.

[45] S. Zhang, Y. Zhan, and D. N. Metaxas, "Deformable segmentation via sparse representation and dictionary learning," *Medical Image Analysis*, vol. 16, no. 7, pp. 1385–1396, 2012.

[46] G. Wang, S. Zhang, H. Xie, D. N. Metaxas, and L. Gu, "A homotopy-based sparse representation for fast and accurate shape prior modeling in liver surgical planning," *Medical image analysis*, vol. 19, no. 1, pp. 176–186, 2015.

[47] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. O'Regan *et al.*, "Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation," *IEEE transactions on medical imaging*, vol. 37, no. 2, pp. 384–395, 2017.

[48] H. Ravishankar, R. Venkataramani, S. Thiruvenkadam, P. Sudhakar, and V. Vaidya, "Learning and incorporating shape models for semantic segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017, pp. 203–211.

[49] A. V. Dalca, J. Guttag, and M. R. Sabuncu, "Anatomical priors in convolutional networks for unsupervised biomedical segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9290–9299.

[50] X. Zhuang, "Challenges and methodologies of fully automatic whole heart segmentation: a review," *Journal of healthcare engineering*, vol. 4, no. 3, pp. 371–407, 2013.

[51] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 2016, pp. 565–571.

[52] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via ibn-net," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 464–479.