Semi-supervised Medical Image Segmentation through Dual-task Consistency

Xiangde Luo[1] Jieneng Chen[3] Tao Song[2] Yinan Chen[2] Guotai Wang[1,*] Shaoting Zhang[1,2]

[1]University of Electronic Science and Technology of China, Chengdu, China
[2] SenseTime Research, Shanghai, China
[3]Tongji University, Shanghai, China
[*] guotai.wang@uestc.edu.cn

Abstract

Deep learning-based semi-supervised learning (SSL) algorithms have led to promising results in medical images segmentation and can alleviate doctors' expensive annotations by leveraging unlabeled data. However, most of the existing SSL algorithms in literature tend to regularize the model training by perturbing networks and/or data. Observing that multi/dual-task learning attends to various levels of information which have inherent prediction perturbation, we ask the question in this work: can we explicitly build task-level regularization rather than implicitly constructing networks- and/or data-level perturbation and then regularization for SSL? To answer this question, we propose a novel dual-task-consistency semi-supervised framework for the first time. Concretely, we use a dual-task deep network that jointly predicts a pixel-wise segmentation map and a geometry-aware level set representation of the target. The level set representation is converted to an approximated segmentation map through a differentiable task transform layer. Simultaneously, we introduce a dual-task consistency regularization between the level set-derived segmentation maps and directly predicted segmentation maps for both labeled and unlabeled data. Extensive experiments on two public datasets show that our method can largely improve the performance by incorporating the unlabeled data. Meanwhile, our framework outperforms the state-of-the-art semi-supervised learning methods. Code is available at: https://github.com/Luoxd1996/DTC

Introduction

Accurate and robust segmentation of organs or lesions from medical images plays an essential role in many clinical applications such as diagnosis and treatment planning (Masood et al. 2015). With a large amount of labeled data, deep learning has achieved the state-of-the-art performance on automatic image segmentation (Long, Shelhamer, and Darrell 2015; Chen et al. 2018). For medical image, however, annotations are often expensive to acquire as both expertise and time are needed to produce accurate annotations, especially in 3D volumetric images. To reduce the labeling cost, recently, many methods are proposed to develop a high-performance model for medical image segmentation with less labeled data. For example, combining user interaction with deep neural network to perform image segmentation interactively can reduce the labeling efforts (Wang et al. 2018a,b). Self-supervised learning approaches utilize unlabeled data to train models in a supervised manner to learn fundamental knowledge for knowledge transfer (Zhu et al. 2020). Semi-supervised learning framework obtains high-quality segmentation results by learning from a limited amount of labeled data and a large set of unlabeled data directly (Li et al. 2020). Weakly supervised learning methods learn from bounding boxes, scribbles or image-level tags for image segmentation rather than using pixel-wise annotation, which reduces the burden of annotation (Dai, He, and Sun 2015; Lin et al. 2016; Lee et al. 2019). In this work, we focus on the semi-supervised segmentation methods, as the performance of weakly supervised learning and self-supervised learning are much limited for segmentation of 3D medical images, and it is more practical to acquire a small set of fully annotated images with a large set of unannotated images.

Many recent successful SSL methods (Yu et al. 2019; Li et al. 2020; Nie et al. 2018; Li, Zhang, and He 2020) incorporate unlabeled data by performing unsupervised consistency regularization. To be specific, they can either add small perturbations to the unlabeled samples and enforce the consistency between the model predictions on the original data and the perturbed data (Yu et al. 2019; Li et al. 2020), or just directly enforce the similar prediction distributions on the entire unlabeled dataset with an adversarial regularization (Nie et al. 2018; Li, Zhang, and He 2020). Thus, we have learned that the essence of the discussed SSL works is to enforce the consistency on predictions of unlabeled data via a regularization term in loss function.

Among the aforementioned SSL works, it is delighted to see Li, Zhang, and He (2020) developed a multi-task network containing the pixel-wise and the shape-aware prediction branches, similar to previous fully supervised works (Wang et al. 2020; Xue et al. 2020). And for SSL, they consider only the shape branch to build the consistent constraints via an adversarial regularization to make prediction distributions on the entire unlabeled dataset be smooth, which still belongs to data-level regularization. We observe that various levels of information from different task branches can complement each other during training, while different focuses can lead to inherent prediction perturbation. For example, if the predictions from pixel-wise branch and shape-aware branch are finally evaluated under the same criterion, we will definitely obtain different results i.e. the prediction perturbations between different tasks. Then we

ask the most significant question in this work: can we explicitly build task-level regularization totally different from previous data-level regularization? Apparently the answer is yes, on the condition that the output of different task branches should be mapped/transformed to the same predefined space, where we are capable to explicitly enforce the consistency regularization between two prediction maps.

To this end, we propose a novel dual-task-consistency model for semi-supervised medical image segmentation. Our main idea is to build the consistency between a global-level level set function regression task and a pixel-wise classification task to take geometric constraints into account and utilize the unlabeled data. Our framework consists of three parts: the first part is dual-task segmentation network. Specifically, we model a segmentation problem as two different representations (tasks): predicting a pixel-wise classification map and obtaining a global-level level set function where the zero level let is the segmentation contour. We use a two-branch network to predict these two representations, and using a CNN to predict level set function is inspired by (Ma et al. 2020; Xue et al. 2020) to embed global information and geometric constraints into a network for better performance. The second part of this framework is a differentiable task transform layer. We use a smooth Heaviside layer (Xue et al. 2020) to convert the level set function to a segmentation probability map in a differentiable way. The third part is a combination loss function for supervised and unsupervised learning, where we design a dual-task-consistency loss function to minimize the difference between the predicted pixels-wise segmentation probability map and the probability map converted from the level set function, which can be used to boost the performance of fully supervised learning and also can be used to utilize the unlabeled data for unsupervised learning efficiently. Our proposed framework has been applied to two different semi-supervised medical image segmentation tasks: left atrium segmentation from MRI and pancreas segmentation from CT. Experimental results indicate that our proposed algorithm can improve segmentation accuracy, compared to other state-of-the-art semi-supervised segmentation methods. Overall, we present a simple yet efficient semi-supervised medical image segmentation method with dual-task consistancy, which leverages the unlabeled data by encouraging consistent predictions of the same input under different tasks. Our findings during experiments include:

1) In the fully supervised setting, our dual-task consistency regularization outperforms the separate and joint supervision of dual tasks.

2) In the semi-supervised setting, the proposed framework outperforms state-of-the-art semi-supervised medical image segmentation frameworks on several clinical datasets.

3) Compared with existing methods, the proposed framework requires less training time and computational cost. Meanwhile, it is directly applicable to any semi-supervised medical image segmentation scene and can easily be extended to use additional tasks given that there exists a differentiable transform between/among tasks.

Related Works

Semi-Supervised Medical Image Segmentation

For semi-supervised medical image segmentation, traditional methods mainly use hand-crafted features to design a model to perform segmentation, which includes the prior-based models (You et al. 2011) and the clustering-based models (Portela, Cavalcanti, and Ren 2014). The performance of the hand-crafted features-based models often relies on the hand-crafted features' representation capacity. For example, the prior-based models need to design the specific prior information for different organs, which can hardly generalize to other organs. The clustering-based models are often parameter-sensitive and not robust enough, which leads to the poor prediction for objects with large shape variance.

With the ability to learn high-level semantic features automatically, deep learning has been widely used for medical image segmentation (Ronneberger, Fischer, and Brox 2015). Recently, almost all semi-supervised medical image segmentation frameworks are based on deep learning. Bai et al. (2017) developed an iterative framework where in each iteration, pseudo labels for unannotated images are predicted by the network and refined by a Conditional Random Field (CRF) (Krähenbühl and Koltun 2011), then the new pseudo labels are used to update the network. Using adversarial learning to utilize the unlabeled data is also a popular way for semi-supervised medical image segmentation. Zhang et al. (2017) proposed a new deep adversarial network (DAN) model for biomedical image segmentation by encouraging the segmentation of unannotated images to be similar to those of the annotated ones. Yu et al. (2019) extended the mean teacher model (Tarvainen and Valpola 2017) with uncertainty map guidance for semi-supervised left atrium segmentation. Li, Zhang, and He (2020) introduced a shape-aware semi-supervised segmentation strategy to leverage the unlabeled data and to enforce a geometric shape constraint on the segmentation output. Differently, our method takes advantage of geometric constraints and dual-task-consistency, which is simple yet effective for semi-supervised medical image segmentation.

Consistency Regularization

The consistency regularization plays a vital role in computer vision and image processing, especially in semi-supervised learning. For examples, Sajjadi, Javanmardi, and Tasdizen (2016) proposed a regularization with stochastic transformations and perturbations for deep semi-supervised learning, and learned from unlabeled images by minimizing the difference between the predictions of multiple passes of a training sample. Tarvainen and Valpola (2017) introduced a teacher-student consistency model to make full use of the unlabeled data, where the student model learns from the teacher model by minimizing the segmentation loss on the labeled data and the consistency loss with respect to the targets from the teacher model on all input data. Jeong et al. (2019) used consistency constraints as a tool for enhancing detection performance by making full use of available unlabeled data. Li et al. (2020) introduced a transformation-consistent based semi-supervised segmentation method, which encour-

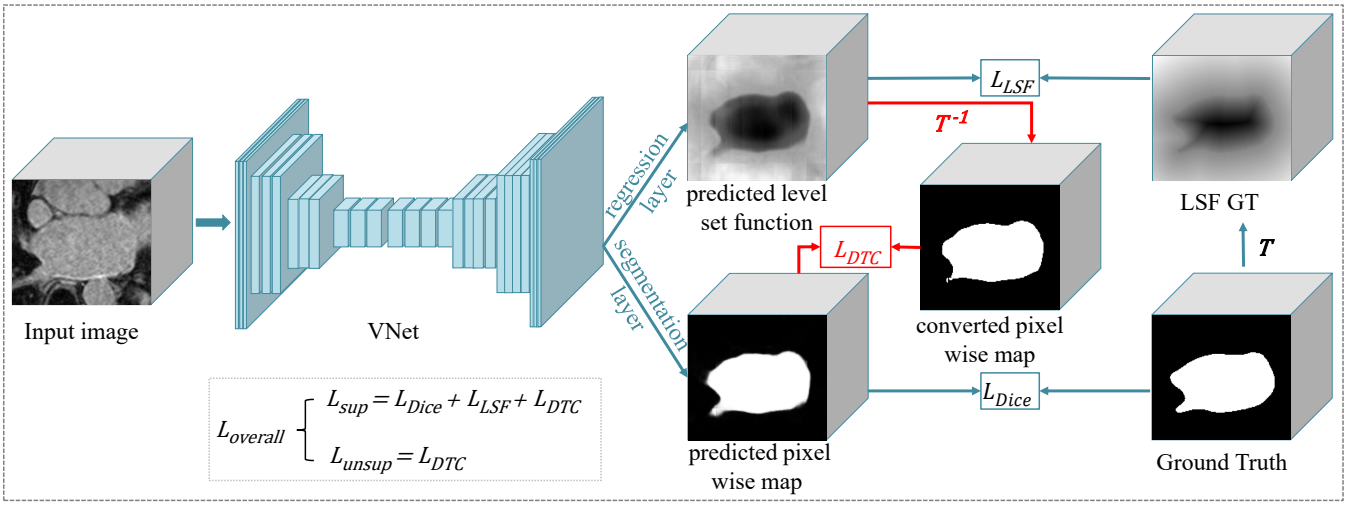

Figure 1: Overview of the proposed dual-task-consistency framework for semi-supervised medical image segmentation. The network consists of a pixel-wise classification head (task1) and a level set function regression head (task2), which employs a widely-used encoder-decoder network as the backbone, i.e., VNet (Milletari, Navab, and Ahmadi 2016). The model is optimized by minimizing supervised losses L_{Dice}, L_{LSF} on labeled data and the dual-task-consistency loss L_{DTC} on both unlabeled data and labeled data. The T function is used to transform the ground truth label map into a level set representation for supervised training. The T^{-1} function converts the level set function to a probability map to calculate the L_{DTC}.

ages consistent predictions of the network-in-training for the same input under different perturbations. However, these works just consider the consistency when the input under different perturbations and transformation, which ignore the consistency of different tasks. In addition, these methods need to perform forward pass two or more times for calculating the consistency loss, which increases the computational cost and running time. More recently, Zamir et al. (2020) utilized the consistency cross different tasks based on inference-path invariance, indicating it is promising to investigate task consistency. The limitation is that they require labeled data in a fully supervised manner and only studied on low-level vision tasks. In contrast to aforementioned methods, our framework aims to utilize the unlabeled data by minimizing the consistency between two tasks of a network, which considers the difference of different tasks and just needs to perform inference once. To the best of our knowledge, our work is the first to construct the task-consistency constraint in a semi-supervised learning fashion to leverage unlabeled data.

Methods

In this section, we introduce our proposed semi-supervised medical image framework based on dual-task-consistency. The overall framework is illustrated in Figure 1, which consists of two heads, the classification head for pixel-wise probability map and the regression head for level set representation of the target. The segmentation network takes a 3D medical image as input, and predicts the level set function and pixel-wise probability map at the same time. As a segmentation result can be represented by both a pixel-level label map and a high-level contour related to a level set function, these two predictions should be consistent for the segmentation task. To utilize the unlabeled data, we propose a novel dual-task-consistency strategy, which learns from unlabeled data by minimizing the difference between the predicted pixel-wise label and the level set function. To build the consistency, a transform layer is used to convert the level set function to a pixel-wise probability map, which is implemented by smooth Heaviside function. In the following two subsections, we first introduce the dual-task consistency strategy, then introduce the semi-supervised training for segmentation through dual-task consistency.

Dual-task Consistency

In general semi-supervised learning, consistency losses are designed to encourage smooth predictions in a data-level, i.e. the predictions of same data under different transformations (Li et al. 2020) and perturbations (Ouali, Hudelot, and Tami 2020) should be the same. In contrast to data-level consistency, we enforce the task-level consistency between the pixel-level classification task, defined as task1 and the level set regression task, defined as task2.

In existing works, pixel-wise classification for segmentation has been widely studied while level-set function (Li et al. 2005) is a traditional task that captures geometric active contours and distance information, which rejuvenates recently when combining with CNN (Wang et al. 2020). We introduce the level set function defined as follows:

$$\mathcal{T}(x) = \begin{cases} -\inf\limits_{y \in \partial S} \|x - y\|_2, & x \in \mathcal{S}_{\text{in}} \\ 0, & x \in \partial \mathcal{S} \\ +\inf\limits_{y \in \partial S} \|x - y\|_2, & x \in \mathcal{S}_{\text{out}} \end{cases} \tag{1}$$

where x, y are two different pixels/voxels in a segmentation

Algorithm 1 Semi-supervised training through Dual-task consistency,

Input: $\mathbf{x}_i \in \mathcal{D}_l + \mathcal{D}_n, \mathbf{y}_i \in \mathcal{D}_l$

Output: Dual-task model's parameter θ_1 for segmentation head, θ_2 for level-set function (LSF) head and θ for shared-weights backbone network

$f_1(x)$ = segmentation task branch with shared parameter θ and segmentation head's parameter θ_1

$f_2(x)$ = LSF task branch with shared parameter θ and LSF head's parameter θ_2

while stopping criterion not met: **do**

  Sample batch $b_l = (\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}_l$ and $b = b_l + b_u$, where $b_u = \mathbf{x}_i \in \mathcal{D}_u$

  Generating LSF ground truth $\mathcal{T}(\mathbf{y}_i)$ according to Equation. 1

  Computing dual-task predictions $f_1(\mathbf{x}_i)$ and $f_2(\mathbf{x}_i)$, $i \in \{1, ..., N\}$ where N denotes the batch size

  Applying task transform layer $\mathcal{T}^{-1}(f_2(\mathbf{x}_i))$ according to Equation. 2

  $\mathcal{L}_{DTC}(\mathbf{x}) = \frac{1}{|b|} \sum_{\mathbf{x}_i \in b} \left\| f_1(\mathbf{x}_i) - \mathcal{T}^{-1}(f_2(\mathbf{x}_i)) \right\|^2$

  $\mathcal{L}_{LSF}(\mathbf{x}, \mathbf{y}) = \frac{1}{|b_l|} \sum_{\mathbf{x}_i, \mathbf{y}_i \in b_l} \left\| f_2(\mathbf{x}_i) - \mathcal{T}(\mathbf{y}_i) \right\|^2$

  $\mathcal{L}_{Seg}(\mathbf{x}, \mathbf{y}) = 1 - \frac{1}{|b_l|} \sum_{\mathbf{x}_i, \mathbf{y}_i \in b_l} 2 \frac{\sum f_1(\mathbf{x}_i) \mathbf{y}_i}{\sum f_1(\mathbf{x}_i) + \sum \mathbf{y}_i}$

  $\mathcal{L}_{total} = \mathcal{L}_{Seg} + \mathcal{L}_{LSF} + \lambda_d \mathcal{L}_{DTC}$

  Computing gradient of loss function $\mathcal{L}_{total}$ and update network parameters θ_1, θ_2 and θ by back propagation.

end while

return θ_1, θ_2 and θ

mask, the ∂S is the zero level set and also represents the contour of the target object. $\mathcal{S}_{\text{in}}$ and $\mathcal{S}_{\text{out}}$ denote the inside region and outside region of the target object. Then we define $\mathcal{T}(x)$ as the task transform from segmentation map to level-set function map in Equation. 1. To map the output of LSF task to the space of segmentation output, it is natural to think of using an inverse transform of $\mathcal{T}(x)$. However, it is impractical to integrate the exact inverse transform of $\mathcal{T}(x)$ in training due to the non-differentiability. Hence, we utilize a smooth approximation to the inverse transform of level-set function, provided that we want to guarantee the values of $\mathcal{S}_{\text{in}}$ are assigned to 1 while those of $\mathcal{S}_{\text{out}}$ are assigned to 0 in the transformed prediction map, which is defined as:

$$\mathcal{T}^{-1}(z) = \frac{1}{1 + e^{-k \cdot z}} = \sigma(k \cdot z) \tag{2}$$

where z means the level set value at pixel/voxel x. The formulation of $\mathcal{T}^{-1}(z)$ is delicate and simple as it is equal to Sigmoid function with the input multiplied by a factor k, which is selected as large as possible to approximate inverse transform of $\mathcal{T}(x)$. Thus, $\mathcal{T}^{-1}(z)$ can easily be implemented as an modified activate function followed by task2's output. Then the differentiability can be proved as follows:

$$\begin{aligned} \frac{\partial \mathcal{T}^{-1}}{\mathrm{d}z} &= \left(\frac{1}{1 + e^{-k \cdot z}} \right)' \\ &= k \cdot \frac{1}{1 + e^{-kz}} \cdot \left(1 - \frac{1}{1 + e^{-kz}} \right) \end{aligned} \tag{3}$$

Though such approximate transform function will map the prediction space of task2 to be the same with that of task1, it naturally introduces a task-level prediction difference since task1 focuses on pixel-level reasoning while task2 attends to geometric structure information. Thus, for input $\mathbf{X}$ from a dataset $\mathcal{D}$, we define the dual-task-consistency loss $\mathcal{L}_{DTC}$ enforcing consistency between task1's prediction $f_1(\mathbf{x}_i)$ and the transformed map of task2's prediction $\mathcal{T}^{-1}(f_2(\mathbf{x}_i))$:

$$\begin{aligned} \mathcal{L}_{DTC}(\mathbf{x}) &= \sum_{\mathbf{x}_i \in \mathcal{D}} \left\| f_1(\mathbf{x}_i) - \mathcal{T}^{-1}(f_2(\mathbf{x}_i)) \right\|^2 \\ &= \sum_{\mathbf{x}_i \in \mathcal{D}} \left\| f_1(\mathbf{x}_i) - \sigma(k \cdot f_2(\mathbf{x}_i)) \right\|^2 \end{aligned} \tag{4}$$

Semi-supervised training through Dual-Task-Consistency: Let $\mathcal{D}_l$ and $\mathcal{D}_u$ be the labeled and unlabeled dataset, respectively. Let $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$ be the whole provided dataset. We denote labeled data pair as $(\mathbf{X}, \mathbf{Y}) \in \mathcal{D}_l$ and unlabeled data as $\mathbf{X} \in \mathcal{D}_u$, where $\mathbf{Y}$ is groundtruth segmentation mask. We denote voxel-level pair as $(x, y) \in (\mathbf{X}, \mathbf{Y})$. For labeled data $\mathcal{D}_l$, we define the supervised loss for segmentation task as commonly used dice loss :

$$\begin{aligned} \mathcal{L}_{Seg}(\mathbf{x}, \mathbf{y}) &= \sum_{\mathbf{x}_i, \mathbf{y}_i \in \mathcal{D}_l} \mathcal{L}_{Dice}(\mathbf{x_i}, \mathbf{y}_i) \\ &= \sum_{\mathbf{x}_i, \mathbf{y}_i \in \mathcal{D}_l} \left(1 - \frac{2 \sum_{x_j \in \mathbf{x}_i, y_j \in \mathbf{y}_i} f_1(x_i) \, y_i}{\sum_{x_j \in \mathbf{x}_i, y_j \in \mathbf{y}_i} f_1(x_j) + \sum_{y_j \in \mathbf{y}_i} y_j}\right) \end{aligned} \tag{5}$$

where the summation for $\sum_{x_j \in \mathbf{x}_i, y_j \in \mathbf{y}_i}$ denotes voxel-wise sum in a 3D image, and the summation for $\sum_{\mathbf{x}_i, \mathbf{y}_i \in \mathcal{D}_l}$ denotes image-level sum in a dataset. Then we define the supervised loss for LSF task as $\mathcal{L}_2$ loss between the predicted probability map $f_2(\mathbf{x})$ and the transformed ground truth map $\mathcal{T}(\mathbf{y})$:

$$\mathcal{L}_{LSF}(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{x}_i, \mathbf{y}_i \in \mathcal{D}_l} \left\| f_2(\mathbf{x}_i) - \mathcal{T}(\mathbf{y}_i) \right\|^2 \tag{6}$$

It is noteworthy that for annotated images, the ground truth level set function for the LSF task can be automatically generated from labeled segmentation mask $\mathbf{Y}$ through aforementioned task transform function $\mathcal{T}$. The final loss is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{Seg} + \mathcal{L}_{LSF} + \lambda_d \mathcal{L}_{DTC} \tag{7}$$

where $\mathcal{L}_{Seg}$ and $\mathcal{L}_{LSF}$ are only used for labeled data, while $\mathcal{L}_{DTC}$ is used for both labeled and unlabeled data during training, and therefore the two tasks can jointly optimize the network with either labeled data or unlabeled data in a semi-supervised fashion. Following (Tarvainen and Valpola 2017; Yu et al. 2019), we use a time-dependent Gaussian warming up function $\lambda_d(t) = e^{(-5(1 - \frac{t}{t_{max}})^2)}$ to control the balance between the supervised loss and unsupervised consistency loss, where t denotes the current training step and t_{max} is the maximum training step. The used training algorithm for semi-supervised segmentation through dual-task consistency is shown in Algorithm. 1.

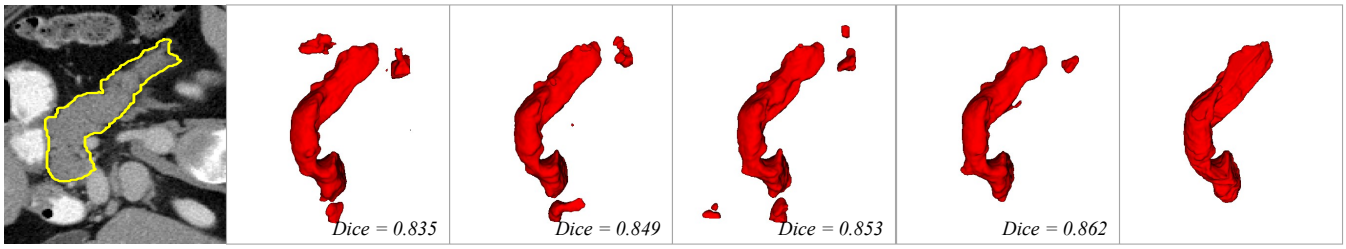

Figure 2: 3D Visualization of different training methods for pancreas segmentation. 12 annotated images without unannotated images were used for training. GT: ground truth. (best viewed in color)

Method	Scans used		Metrics				Cost	
	Labeled	Unlabeled	Dice (%)	Jaccard (%)	ASD (voxel)	95HD (voxel)	Params (M)	Training time (h)
Seg	12	0	70.63	56.72	6.29	22.54	9.44	**2.1**
LSF	12	0	71.78	57.55	6.31	20.74	9.44	2.1
Seg + LSF	12	0	73.08	58.65	4.47	18.04	9.44	2.2
Seg + LSF + DTC	12	0	**74.84**	**60.78**	**2.17**	**9.34**	9.44	2.3
Seg	62	0	81.78	69.65	1.34	5.13	9.44	**2.3**
LSF	62	0	82.25	70.23	**1.18**	5.19	9.44	2.5
Seg + LSF	62	0	82.46	70.61	1.22	4.97	9.44	2.5
Seg + LSF + DTC	62	0	**82.80**	**71.05**	1.45	**4.67**	9.44	2.5

Table 1: Ablation study of our dual-task consistency method on the Pancreas CT dataset.

Experiments and Results

Datasets and Pre-processing: To evaluate the proposed method, we apply our algorithm on two different datasets. The first is left atrial dataset[1], which consists of 100 3D gadolinium-enhanced MR images, with a resolution of $0.625 \times 0.625 \times 0.625mm$. Following (Yu et al. 2019; Li, Zhang, and He 2020), we use 80 scans for training and 20 scans for validation, and apply the same pre-processing methods. The second is pancreas dataset[2], which includes 82 abdomen CT images. Following (Xia et al. 2020), we randomly split them into 62 images for training and 20 images for testing. In pre-processing, we use the soft tissue CT window range of $[-125, 275]$ HU (Zhou et al. 2019), and resample all images to an isotropic resolution of $1.0 \times 1.0 \times 1.0mm$. Finally, we crop the images centering at the pancreas region based on the ground truth with enlarged margins (25 voxels) and normalize them as zero mean and unit variance. In this work, we report the performance of all methods trained with 20% labeled images and 80% unlabeled images, which is the typical semi-supervised learning experimental setting (Xia et al. 2020; Yu et al. 2019; Li, Zhang, and He 2020).

Implementation Details and Evaluation Metrics: We implement our framework in PyTorch (Paszke et al. 2019), using an NVIDIA 1080TI GPU. In this work, we use VNet (Milletari, Navab, and Ahmadi 2016) as the backbone for all experiments, and we implement dual-task VNet by adding a new regression layer at the end of the original VNet. The framework is trained by an SGD optimizer for 6000 iterations, with an initial learning rate (lr) 0.01 decayed by 0.1 every 2500 iterations. The batch size is 4, consisting of 2 labeled images and 2 unlabeled images. Following (Xue et al. 2020), the value of k is set to 1500 in this work. We randomly crop $112 \times 112 \times 80$ (3D MRI Left Atrium) and $96 \times 96 \times 96$ (3D CT Pancreas) sub-volume as the network input. To avoid over-fitting, we use the standard on-the-fly data augmentation methods during training stage (Yu et al. 2019). Note that, in this work, the level set function is generated before the training phase rather on-the-fly, since the level set function is transform-invariant, which in result significantly speed up the training procedure. In the inference phase, we use a sliding window strategy to obtain the final results, which with a stride of $18 \times 18 \times 4$ for left atrium and $16 \times 16 \times 16$ for pancreas. At the inference time, we use the output of pixel-wise classification branch as the segmentation result. For a fair comparison, we do not use any post-processing or ensemble methods. Following (Yu et al. 2019), we use four metrics to quantitatively evaluate our method, including Dice, Jaccard, the average surface distance (ASD), and the 95% Hausdorff Distance (95HD).

The Effects of Different Tasks: To investigate the individual impact of different tasks, we first only use labeled images for training and analyze how the dual-task consistency performs when only labeled images are used. We trained the network for pancreas segmentation using the 12 labeled data and all the 62 labeled data, respectively. We compared different training strategies: 1) only using the branch for task1 (Seg), 2) only using the branch for task 2 (LSF), 3) using the two branches for task1 and task2 simultaneously (Seg + LSF), and 4) and our proposed dual-task consistency method

[1]http://atriaseg2018.cardiacatlas.org

[2]https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT

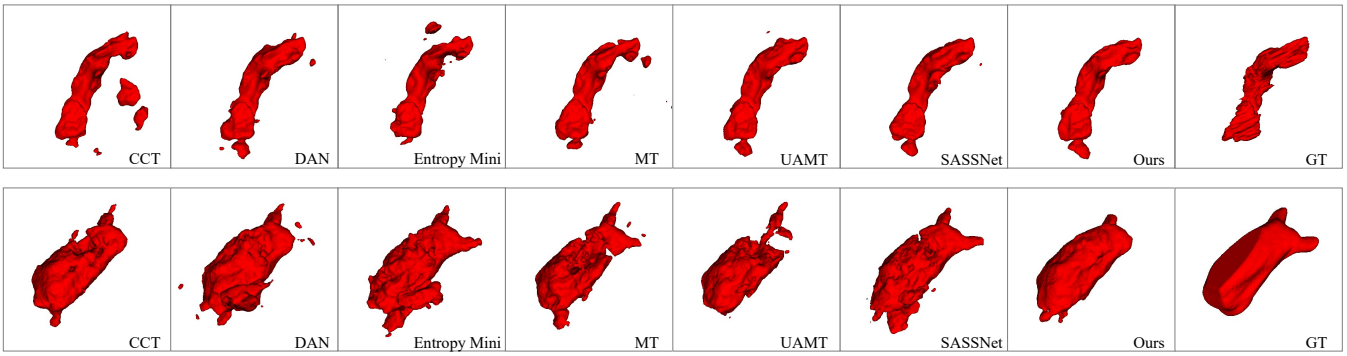

Figure 3: 3D Visualization of different semi-supervised segmentation methods under 20% labeled data (best viewed in color). The first row is a pancreas segmentation result and second row is a left atrium segmentation result.

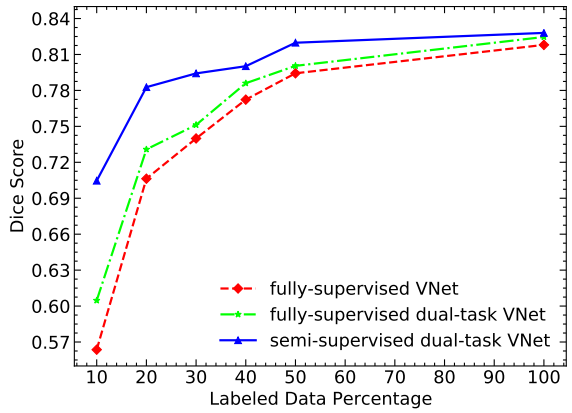

Figure 4: The pancreas segmentation performance of our semi-supervised approach with different ratio of labeled data. The dashed red and lime curves show performance of fully-supervised VNet and dual-task VNet respectively, where they were trained with only the available labeled data.

(Seg + LSF + DTC). The performance of these variants is listed in Table. 1. It shows that the level set function regression is helpful for medical image segmentation. It also can be observed that dual-task consistency consistently improves the performance of the dual-task VNet on 12 labeled scans and 62 labeled scans. Figure. 2 shows some visualization of different training methods, which further show the superiority of our proposed dual-task-consistency.

Effectiveness of Dual-task-Consistency for Semi-supervised Learning: Secondly, we performed a study on data utilization efficiency of our approach compared to the fully supervised VNet and dual-task VNet that only use available annotated images for training on Pancreas CT dataset. We draw the Dice score of the results in Figure.4. It can be observed that the semi-supervised method consistently performs better than the supervised approach in different labeled data settings, demonstrating that our method effectively utilizes the unlabeled data and brings performance gains. It also can be found that the performance gap between fully supervised method and semi-supervised approach narrows as more labeled images are available, which conforms to the common sense. When the number of labeled data is small, our method also can obtain a better segmentation result than fully supervised method, indicating the promising potential of our proposed approach for further clinical use.

Comparison with Other Semi-supervised Methods: We compared our framework with six state-of-the-art semi-supervised segmentation methods, including deep adversarial network (DAN) (Zhang et al. 2017), entropy minimization approach (Entropy Mini) (Vu et al. 2019), cross-consistency training method (CCT) (Ouali, Hudelot, and Tami 2020), mean teacher self-ensembling model (MT) (Tarvainen and Valpola 2017), uncertainty-aware mean teacher model(UA-MT) (Yu et al. 2019) and shape-aware adversarial network (SASSNet) (Li, Zhang, and He 2020). Note that we used the official code and results of DAN, MT, UA-MT, SASSNet, and reimplemented the Entropy Mini and CCT for medical image segmentation, since the limitation of GPU memory, we used one main decoder and three auxiliary decoders as CCT's implementation.

We first evaluate our proposed framework on Pancreas CT. Table. 2 shows the quantitative comparison of these methods. Compared with fully supervised VNet trained with only 12 annotated images, all semi-supervised methods taking advantages of unannotated images improve the segmentation performance significantly. The MT, UA-MT and CCT achieve slightly better performance than Entropy Mini and DAN, demonstrating that perturbation-based consistency loss is helpful for the semi-supervised segmentation problem. Moreover, the UA-MT is better than MT, since the uncertainty map can guide the student model learning efficiently. The SASSNet achieves the top performance among the existing methods, indicating the shape prior is useful for semi-supervised image segmentation. Notably, our framework achieves better performance than the state-of-the-art semi-supervised methods on all the evaluation metrics without using a complex multiple network architecture, corroborating that our dual-task-consistency has the full capability to draw out the rich information from the unlabeled data. Meanwhile, our framework does not require any multiple inference or iteratively update scheme, which reduces the computational memory cost and running time.

Table 2: Quantitative comparison between our methods and other semi-supervised methods on the Pancreas CT dataset. The first and second row are our fully supervised baseline, the last row is our proposed method, others are previous methods.

Method	Scans used		Metrics				Cost	
	Labeled	Unlabeled	Dice (%)	Jaccard (%)	ASD (voxel)	95HD (voxel)	Params (M)	Training time (h)
VNet	12	0	70.63	56.72	6.29	22.54	9.44	**2.1**
VNet	62	0	81.78	69.65	1.34	5.13	9.44	2.3
MT (NeurIPS'17)	12	50	75.85	61.98	3.40	12.59	9.44	2.9
DAN (MICCAI'17)	12	50	76.74	63.29	2.97	11.13	12.09	3.3
Entropy Mini (CVPR'19)	12	50	75.31	61.73	3.88	11.72	9.44	**2.2**
UA-MT (MICCAI'19)	12	50	77.26	63.82	3.06	11.90	9.44	3.9
CCT (CVPR'20)	12	50	76.58	62.76	3.69	12.92	15.65	4.1
SASSNet (MICCAI'20)	12	50	77.66	64.08	3.05	10.93	20.46	3.9
Ours	12	50	**78.27**	**64.75**	**2.25**	**8.36**	**9.44**	2.5

Table 3: Quantitative comparison between our methods and other semi-supervised methods on the Left Atrium MRI dataset. The first and second row are our fully supervised baseline, the last row is our proposed method, others are previous methods.

Method	Scans used		Metrics				Cost	
	Labeled	Unlabeled	Dice (%)	Jaccard (%)	ASD (voxel)	95HD (voxel)	Params (M)	Training time (h)
VNet	16	0	86.03	73.26	5.75	17.93	9.44	**1.8**
VNet	80	0	91.14	83.32	1.52	5.75	9.44	2.0
MT(NeurIPS'17)	16	64	88.23	79.29	2.73	10.64	9.44	3.2
DAN (MICCAI'17)	16	64	87.52	78.29	2.42	9.01	12.09	3.7
Entropy Mini (CVPR'19)	16	64	88.45	79.51	3.72	14.14	9.44	**1.9**
UA-MT (MICCAI'19)	16	64	88.88	80.21	2.26	7.32	9.44	3.6
CCT (CVPR'20)	16	64	88.83	80.06	2.49	8.44	15.65	3.9
SASSNet (MICCAI'20)	16	64	89.27	80.82	3.13	8.83	20.46	4.4
Ours	16	64	**89.42**	**80.98**	**2.10**	**7.32**	**9.44**	2.2

We further validate our proposed method on Left Atrium MRI data, which is a widely-used dataset for semi-supervised medical image segmentation (Yu et al. 2019; Li, Zhang, and He 2020). A quantitative comparison of these methods is shown in Tabel. 3. It can be found that our method achieved the best accuracy than other methods on all the evaluation metrics, especially in term of ASD and 95HD. Figure. 3 shows some visualization of pancreas segmentation and left atrium segmentation. Compared with other methods, our results have higher overlap ratio with the ground truth and produce less false positives and preserve more details, which further indicates the effectiveness, generalization and robustness of our proposed method. Furthermore, we investigated the training cost of different approaches. The quantitative comparison of network's parameters and training time are listed in Table.2 and Table.3. It can be observed that, our framework requires less training time than MT, DAN, UAMT, CCT and SASSNet, since our framework use a simple network with fewer parameters and does not need to pass an image many times in an iteration. Compared with Entropy Mini and fully supervised baseline, our method achieved better accuracy with comparable computational cost. Thus, our experiments prove that our method attains the best accuracy, networks' parameters and computational-cost trade-offs.

Discussion and Conclusion

In this paper, we have presented a novel and simple semi-supervised medical image segmentation framework through dual-task consistency, which is a task-level consistency-based framework for semi-supervised segmentation. We use a dual-task network that simultaneously predicts a pixel-level classification map and a level set representation of the segmentation that is able to capture global-level shape and geometric information. In order to build a semi-supervised training framework, we enforce dual-task consistency between classification map prediction and LSF prediction via a task-transform layer. We achieve stat-of-the-art results on two 3D medical image datasets including left atrial dataset in MR scans and pancreas dataset in CT scans. The superior performance demonstrates the effectiveness, robustness and generalization of our proposed framework. In this work, we focus on single-class segmentation to simplify the presentation. However, our method extends to the multi-class case in a straightforward manner.

In addition, our proposed method can easily be extended to use additional tasks such as edge extraction (Zhen et al. 2020) and key-points estimation (Cheng et al. 2020) as long as there exists differentiable transform between two tasks. We also hope to inspire the whole computer vision community, as it is possible to construct tasks consistency in a semi-supervised fashion in many directions such as two-stream video recognition (Simonyan and Zisserman 2014), multi-task image reconstruction (Zamir et al. 2018, 2020) etc. to leverage a large amount of unlabeled data. In the future, we will extend this method to more computer vision applications to reduce labeling efforts and further investigate the fusion strategy to ensemble all different tasks' prediction results for better performance.

References

Bai, W.; Oktay, O.; Sinclair, M.; Suzuki, H.; Rajchl, M.; Tarroni, G.; Glocker, B.; King, A.; Matthews, P. M.; and Rueckert, D. 2017. Semi-supervised learning for network-based cardiac MR image segmentation. In *MICCAI*, 253–260. Springer. (document)

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 801–818. (document)

Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T. S.; and Zhang, L. 2020. HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation. In *CVPR*, 5386–5395. (document)

Dai, J.; He, K.; and Sun, J. 2015. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *CVPR*, 1635–1643. (document)

Jeong, J.; Lee, S.; Kim, J.; and Kwak, N. 2019. Consistency-based semi-supervised learning for object detection. In *NeurIPS*, 10759–10768. (document)

Krähenbühl, P.; and Koltun, V. 2011. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, 109–117. (document)

Lee, J.; Kim, E.; Lee, S.; Lee, J.; and Yoon, S. 2019. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *CVPR*, 5267–5276. (document)

Li, C.; Xu, C.; Gui, C.; and Fox, M. D. 2005. Level set evolution without re-initialization: a new variational formulation. In *CVPR*, volume 1, 430–436. IEEE. (document)

Li, S.; Zhang, C.; and He, X. 2020. Shape-aware Semi-supervised 3D Semantic Segmentation for Medical Images. In *MICCAI*, 605–613. Springer. (document)

Li, X.; Yu, L.; Chen, H.; Fu, C.-W.; Xing, L.; and Heng, P.-A. 2020. Transformation-Consistent Self-Ensembling Model for Semisupervised Medical Image Segmentation. *IEEE Transactions on Neural Networks and Learning Systems* . (document)

Lin, D.; Dai, J.; Jia, J.; He, K.; and Sun, J. 2016. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR*, 3159–3167. (document)

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*, 3431–3440. (document)

Ma, J.; Wei, Z.; Zhang, Y.; Wang, Y.; Lv, R.; Zhu, C.; Chen, G.; Liu, J.; Peng, C.; Wang, L.; et al. 2020. How Distance Transform Maps Boost Segmentation CNNs: An Empirical Study. In *MIDL*. (document)

Masood, S.; Sharif, M.; Masood, A.; Yasmin, M.; and Raza, M. 2015. A survey on medical image segmentation. *Current Medical Imaging Reviews* 11(1): 3–14. (document)

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 565–571. IEEE. 1, (document)

Nie, D.; Gao, Y.; Wang, L.; and Shen, D. 2018. ASDNet: Attention based semi-supervised deep networks for medical image segmentation. In *MICCAI*, 370–378. Springer. (document)

Ouali, Y.; Hudelot, C.; and Tami, M. 2020. Semi-Supervised Semantic Segmentation with Cross-Consistency Training. In *CVPR*, 12674–12684. (document)

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 8026–8037. (document)

Portela, N. M.; Cavalcanti, G. D.; and Ren, T. I. 2014. Semi-supervised clustering for MR brain image segmentation. *Expert Systems with Applications* 41(4): 1492–1497. (document)

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241. Springer. (document)

Sajjadi, M.; Javanmardi, M.; and Tasdizen, T. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *NeurIPS*, 1163–1171. (document)

Simonyan, K.; and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 568–576. (document)

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 1195–1204. (document)

Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; and Pérez, P. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2517–2526. (document)

Wang, G.; Li, W.; Zuluaga, M. A.; Pratt, R.; Patel, P. A.; Aertsen, M.; Doel, T.; David, A. L.; Deprest, J.; Ourselin, S.; et al. 2018a. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE Transactions on Medical Imaging* 37(7): 1562–1573. (document)

Wang, G.; Zuluaga, M. A.; Li, W.; Pratt, R.; Patel, P. A.; Aertsen, M.; Doel, T.; David, A. L.; Deprest, J.; Ourselin, S.; et al. 2018b. DeepIGeoS: a deep interactive geodesic framework for medical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(7): 1559–1572. (document)

Wang, Y.; Wei, X.; Liu, F.; Chen, J.; Zhou, Y.; Shen, W.; Fishman, E. K.; and Yuille, A. L. 2020. Deep distance transform for tubular structure segmentation in ct scans. In *CVPR*, 3833–3842. (document)

Xia, Y.; Liu, F.; Yang, D.; Cai, J.; Yu, L.; Zhu, Z.; Xu, D.; Yuille, A.; and Roth, H. 2020. 3d semi-supervised learning with uncertainty-aware multi-view co-training. In *WACV*, 3646–3655. (document)

Xue, Y.; Tang, H.; Qiao, Z.; Gong, G.; Yin, Y.; Qian, Z.; Huang, C.; Fan, W.; and Huang, X. 2020. Shape-Aware Or-

gan Segmentation by Predicting Signed Distance Maps. In *AAAI*. (document)

You, X.; Peng, Q.; Yuan, Y.; Cheung, Y.-m.; and Lei, J. 2011. Segmentation of retinal blood vessels using the radial projection and semi-supervised approach. *Pattern recognition* 44(10-11): 2314–2324. (document)

Yu, L.; Wang, S.; Li, X.; Fu, C.-W.; and Heng, P.-A. 2019. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In *MICCAI*, 605–613. Springer. (document)

Zamir, A. R.; Sax, A.; Cheerla, N.; Suri, R.; Cao, Z.; Malik, J.; and Guibas, L. J. 2020. Robust Learning Through Cross-Task Consistency. In *CVPR*, 11197–11206. (document)

Zamir, A. R.; Sax, A.; Shen, W.; Guibas, L. J.; Malik, J.; and Savarese, S. 2018. Taskonomy: Disentangling task transfer learning. In *CVPR*, 3712–3722. (document)

Zhang, Y.; Yang, L.; Chen, J.; Fredericksen, M.; Hughes, D. P.; and Chen, D. Z. 2017. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *MICCAI*, 408–416. Springer. (document)

Zhen, M.; Wang, J.; Zhou, L.; Li, S.; Shen, T.; Shang, J.; Fang, T.; and Quan, L. 2020. Joint Semantic Segmentation and Boundary Detection using Iterative Pyramid Contexts. In *CVPR*, 13666–13675. (document)

Zhou, Y.; Li, Z.; Bai, S.; Wang, C.; Chen, X.; Han, M.; Fishman, E.; and Yuille, A. L. 2019. Prior-aware neural network for partially-supervised multi-organ segmentation. In *ICCV*, 10672–10681. (document)

Zhu, J.; Li, Y.; Hu, Y.; Ma, K.; Zhou, S. K.; and Zheng, Y. 2020. Rubik's Cube+: A Self-supervised Feature Learning Framework for 3D Medical Image Analysis. *Medical Image Analysis* 101746. (document)