Medical Image Analysis 80 (2022) 102517

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency

Xiangde Luo[a,b], Guotai Wang[a,b,*], Wenjun Liao[c], Jieneng Chen[d], Tao Song[e], Yinan Chen[e,f], Shichuan Zhang[g], Dimitris N. Metaxas[h], Shaoting Zhang[a,b]

[a] School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China
[b] Shanghai AI Lab, Shanghai, China
[c] Department of Radiation Oncology, Nanfang Hospital, Southern Medical University, Guangzhou, China
[d] College of Electronics and Information Engineering, Tongji University, Shanghai, China
[e] SenseTime Research, Shanghai, China
[f] West China Biomedical Big Data Center, Sichuan University West China Hospital, Chengdu, China
[g] Department of Radiation Oncology, Sichuan Cancer Hospital and Institute, University of Electronic Science and Technology of China, Chengdu, China
[h] Department of Computer Science, Rutgers University, Piscataway NJ 08854, USA

ARTICLE INFO

Article history:
Received 2 November 2021
Revised 26 May 2022
Accepted 10 June 2022
Available online 15 June 2022

Keywords:
Semi-supervised learning
Uncertainty rectifying
Pyramid consistency
Image segmentation

ABSTRACT

Despite that Convolutional Neural Networks (CNNs) have achieved promising performance in many medical image segmentation tasks, they rely on a large set of labeled images for training, which is expensive and time-consuming to acquire. Semi-supervised learning has shown the potential to alleviate this challenge by learning from a large set of unlabeled images and limited labeled samples. In this work, we present a simple yet efficient consistency regularization approach for semi-supervised medical image segmentation, called Uncertainty Rectified Pyramid Consistency (URPC). Inspired by the pyramid feature network, we chose a pyramid-prediction network that obtains a set of segmentation predictions at different scales. For semi-supervised learning, URPC learns from unlabeled data by minimizing the discrepancy between each of the pyramid predictions and their average. We further present multi-scale uncertainty rectification to boost the pyramid consistency regularization, where the rectification seeks to temper the consistency loss at outlier pixels that may have substantially different predictions than the average, potentially due to upsampling errors or lack of enough labeled data. Experiments on two public datasets and an in-house clinical dataset showed that: 1) URPC can achieve large performance improvement by utilizing unlabeled data and 2) Compared with five existing semi-supervised methods, URPC achieved better or comparable results with a simpler pipeline. Furthermore, we build a semi-supervised medical image segmentation codebase to boost research on this topic: https://github.com/HiLab-git/SSL4MIS.

1. Introduction

Image segmentation is a fundamental and essential task in medical image analysis, especially in image-guided intervention and radiation therapy (Masood et al., 2015; Wang et al., 2018). Recently, with the development of deep learning, Convolutional Neural Networks (CNNs) have achieved state-of-the-art results in many automatic image segmentation tasks (Ronneberger et al., 2015; Long et al., 2015). However, fully supervised learning methods require large and carefully annotated data to train models for good performance. As we know that obtaining a large dataset with pixel-wise annotation is expensive and time-consuming, especially in medical images where the annotation requires medical expertise and clinical experience. How to learn from limited annotations to achieve promising results is becoming a very hot topic in the medical computing community.

Recently, there are many works that attempt to train powerful models with few annotated images. These works can be roughly summarized as (1) Weakly-supervised learning (Valvano et al., 2021), using sparse annotations (bounding boxes, scribbles, image tags) to train models, where the sparse annotations are easier to collect compared with dense annotations; (2) Human-in-the-loop (Wang et al., 2018; Luo et al., 2021c), integrating user interactions with deep learning algorithms to achieve good performance with few pixel-level annotations and user interactions; (3) Semi-/Self-supervised learning (Luo et al., 2021a; Chaitanya et al., 2020),

* Corresponding author.
E-mail addresses: guotai.wang@uestc.edu.cn (G. Wang), Rutgers.shaoting@gmail.com (S. Zhang).

https://doi.org/10.1016/j.media.2022.102517
1361-8415/

that utilizes limited annotated images and extensive unlabeled images to train deep networks with high performance with low annotation budget. In this work, we focus on semi-supervised learning as it is closer to real clinical scenarios and can largely reduce the workload of annotators for the development of deep learning models.

Semi-supervised learning aims to achieve promising results by combining a few labeled data and many unlabeled data, so the key step is to design efficient supervision for unlabeled data. As a result, many methods have been presented to effectively leverage unannotated images. These methods can be mainly categorized into two types: (1) Pseudo-label-based iterative learning strategy (Lee et al., 2013), this approach firstly trains a model on the labeled data and generates pseudo labels for the unlabeled data then refines these pseudo labels, finally uses these refined pseudo labels to retrain a model and repeats this strategy several times to update these pseudo labels and the segmentation model, iteratively. (2) Consistency-based joint training (Tarvainen and Valpola, 2017), which learns from labeled data and unlabeled data in a unified framework, where supervised loss and consistency loss are used for learning from labeled and unlabeled data, respectively. With the success of these strategies in general machine learning, many works have extended them for semi-supervised medical image computing, including classification and segmentation. Bai et al (Bai et al., 2017) presented pseudo-label-based iterative learning for semi-supervised cardiac structure segmentation from MRI, where all pseudo labels are refined by CRF for further model updating. Mean-teacher model (Tarvainen and Valpola, 2017) and its extensions (Cui et al., 2019; Yu et al., 2019; Wang et al., 2020b; Hang et al., 2020) also achieved increasing attention in semi-supervised medical image segmentation, where these methods learn by minimizing the discrepancy between the output of a teacher model and that of a student model. Other methods used some recent techniques also achieved surprising results, such as deep adversarial network (Zheng et al., 2019), cross task consistency (Luo et al., 2021a) and attention mechanism (Nie et al., 2018).

Differently from the above works, we explore utilizing the unlabeled data in a simpler yet more efficient way. We propose uncertainty rectified pyramid consistency for semi-supervised medical image segmentation tasks. First, inspired by pyramid network (Lin et al., 2017) and deep supervision (Lee et al., 2015) for fully supervised learning, we use a network to obtain a pyramid of predictions at multiple scales. Second, to leverage images without labels, we propose pyramid consistency with the assumption that the predictions of the same object at different scales should be close to each other. Considering that there is a lack of labeled data and the up-sampling processing may lead to inaccurate predictions, encouraging them to be consistent at each pixel directly may be affected by outliers and lead to a performance drop (Yu et al., 2019; Cao et al., 2020; Xia et al., 2020). To alleviate this problem, many works used uncertainty maps to filter the unreliable pixels and showed promising results (Cao et al., 2020; Yu et al., 2019; Zheng and Yang, 2021). However, previous methods estimated the uncertainty of each target prediction with Monte Carlo sampling (Gal and Ghahramani, 2016), which needs huge computational costs as it requires multiple forward passes to obtain the uncertainty in each iteration. In this work, we estimate the uncertainty via the prediction discrepancy among multi-scale predictions, which just needs a single forward pass. Afterwards, we use the pixel-wise prediction uncertainty to weigh the pyramid consistency regularization, where pixels with higher uncertainty are assigned with a lower weight. Based on the uncertainty estimation, we further introduce an uncertainty minimization regularization during the training stage to encourage the model to become more confident. The results show that the pyramid consistency, uncertainty rectification module and uncertainty minimization constraint can boost networks to learn from unlabeled data. In addition, the uncertainty rectification module leads to large and moderate performance improvements compared with just using uncertainty minimization constraint and pyramid consistency, respectively. Therefore, the Uncertainty Rectified Pyramid Consistency (URPC) can be used to train models for semi-supervised medical image segmentation in an end-to-end manner. The main contributions of this article are as follows:

1) We present a simple yet efficient semi-supervised method for medical image segmentation by combining pyramid consistency and uncertainty rectifying. To the best of our knowledge, this is the first attempt to directly use the pyramid consistency in a single network for semi-supervised learning.
2) We introduce a single forward pass-based uncertainty estimation method by measuring multi-scale discrepancy and further integrate it into the pyramid consistency framework to more reliably learn from unlabeled data.
3) Experiments on two public datasets and one in-house dataset for lesion and organ segmentation demonstrate the effectiveness of the proposed semi-supervised methods. In addition, we release all implementation of this work and provide several examples on public datasets. It may bring some potential benefits for semi-supervised medical image segmentation research.

This work extends from our previous work published in MICCAI-2021 (Luo et al., 2021b). In this extension, a more comprehensive literature review, related works, experiment descriptions, and discussion are provided. We further provide deeper analyses of the proposed methods, especially exploring how to utilize the multi-scale information for semi-supervised learning. Then, we evaluate our method on two public datasets, including pancreas segmentation from CT and whole-brain tumor segmentation from MRI and one in-house clinical dataset for nasopharyngeal carcinoma segmentation. Finally, we build a semi-supervised medical image segmentation benchmark and re-implemented several recent methods to promote further semi-supervised learning research in the future.

2. Related works

2.1. Medical image segmentation

Deep learning-based methods have achieved promising results in many image segmentation tasks (Long et al., 2015; Ronneberger et al., 2015). For medical image segmentation, UNet (Ronneberger et al., 2015) and its extensions have been widely used as baselines for further study, especially the famous and powerful nnUNet (Isensee et al., 2021). These extensions mainly focus on data augmentation/processing, network architecture and loss function design. Data augmentation/processing is a simple yet efficient technique to improve model performance and robustness, and its importance has been proved by recent works (Isensee et al., 2021; Xu et al., 2020). Network architecture is also an important component for deep learning-based medical image segmentation algorithms. For example, VNet (Milletari et al., 2016) extended UNet (Çiçek et al., 2016) with residual connections for 3D volumetric medical image segmentation. UNet++ (Zhou et al., 2019b) and UNet3+ (Huang et al., 2020) re-designed the skip connections to aggregate features with different stages and scales to further improve the model performance. After that, the attention mechanism was introduced to calibrate and enhance the feature in the channel and spatial dimensions for better feature representation capacity. Attention UNet (Schlemper et al., 2019) integrated the attention gate into UNet to calibrate skip connected low-level fea-

tures. Roy et al. (2018) proposed a concurrent spatial and channel module ("squeeze-and-excitation" (Woo et al., 2018)) to enhance segmentation networks' performance at the same time. CA-Net (Gu et al., 2020) combined the channel, spatial and scale attention with segmentation networks for explainable medical image segmentation. More recently, transformer (Carion et al., 2020) was used to explicitly model long-range dependency to capture the relation of multi-organs and improve segmentation results (Li et al., 2021). Loss functions aim to minimize the discrepancy between the network predictions and the ground truth, playing an im-replaceable role in model training (Ma et al., 2021). These losses can be mainly summarized as: (1) Distribution-based losses, maximizing the similarity between the prediction and the ground label in distribution space, like cross-entropy loss (Ronneberger et al., 2015); (2) Region-based losses, aiming to minimize the discrepancy between predictions and ground truths, like Dice loss (Milletari et al., 2016), and active contour model-based loss (Chen et al., 2019b; 2021); (3) Distance-based loss, minimizing the boundary distance between predictions and ground truths in euclidean distance space, such as hausdorff distance loss (Karimi and Salcudean, 2019) and boundary loss (Kervadec et al., 2021).

2.2. Semi-supervised learning

2.2.1. Semi-supervised classification

Semi-supervised classification was widely studied in the machine learning community (Chapelle et al., 2009). Entropy minimization (Grandvalet et al., 2005), an extremely simple yet efficient method, proved that minimizing the prediction's entropy on the unlabeled data can improve the model performance and also inspired many following works (Vu et al., 2019). Pseudo-labels (Lee et al., 2013; Wang et al., 2021) trained an initial model on the labeled data and inferred with the unlabeled data to generate the pseudo label and used the pseudo label for further training iteratively. Consistency regularization training, the most popular semi-supervised learning strategy in the deep learning area, applied consistency on perturbed/augmented inputs to encourage the model to produce similar output/distributions for the perturbed/augmented inputs, like temporal ensembling (Samuli and Timo, 2017), mean-teacher (Tarvainen and Valpola, 2017) and their extensions. Other works included imposing consistency and generating the pseudo label through multiple data augmentation strategies also brings performance gain for semi-supervised image classification (Berthelot et al., 2019; Sohn et al., 2020).

2.2.2. Semi-supervised medical image segmentation

Recently, many works have attempted to reduce the cost of pixel-level annotation required by segmentation tasks by using semi-supervised learning. The most popular way is to extended the mean-teacher framework (Tarvainen and Valpola, 2017) to different aspects, such as pixel-wise consistency (Cui et al., 2019), uncertainty calibration (Yu et al., 2019; Cao et al., 2020; Wang et al., 2020b) and transformation-consistent (Li et al., 2020b). Co-training (Qiao et al., 2018; Luo et al., 2022) with uncertainty calibration (Xia et al., 2020) and multi-planar training (Zhou et al., 2019a) also achieved good results in many semi-supervised medical image segmentation tasks. Deep adversarial training (Li et al., 2020a; Zheng et al., 2019) utilized the unlabeled data by using a discriminator to align the distributions of labeled data and unlabeled data. Cross-task consistency encouraged different tasks to achieve a similar representation in the predefined space, such as the segmentation and the size regression (Kervadec et al., 2019), the image reconstruction (Chen et al., 2019a) and the level set regression (Luo et al., 2021a). Differently from existing methods, we utilize the pyramid consistency and uncertainty rectifying in a single model for semi-supervised medical image segmentation, which is very simple yet efficient.

3. Methods

The proposed semi-supervised learning method via Uncertainty Rectified Pyramid Consistency (URPC) is depicted in Fig. 1. In this method, the pyramid prediction network was used to segment images and produce a set of predictions at different scales, i.e., pyramid. For labeled data, the standard supervision loss was used to train the segmentation network. In addition, the network is further regularized by the pyramid consistency to leverage the unlabeled data. We further present an uncertainty rectification to temper the consistency loss at outlier pixels which may have substantially different predictions than the average, potentially due to up-sampling errors or lack of labeled data. Thanks to the pyramid predictions, the uncertainty can be estimated by measuring the discrepancy between these predictions and requires a single forward pass.

To describe this work easily and precisely, we first introduce some default formulations of semi-supervised learning. The training set includes two subsets: labeled data set D_N^l with N annotated samples and unlabeled data set D_M^u with M unannotated images, so the entire train set is $D_{N+M} = D_N^l \cup D_M^u$. Assuming that an image $x_i \in D_N^l$, its ground truth y_i is provided. However, if $x_i \in D_M^u$, its ground truth is not available. $f_\phi(.)$ is used to represent segmentation model with parameter set ϕ.

3.1. Pyramid prediction network for semi-Supervised segmentation

Unlike existing works (Dou et al., 2017; Isensee et al., 2021; Lin et al., 2017) that use multi-scale prediction to accelerate the optimization process and boost the performance in a fully supervised setting, we propose to use the multi-scale information for semi-supervised segmentation in this work. We first extend the vanilla segmentation network to a pyramid prediction network ($f_\phi(.)$) that generates a set of segmentation at different scales, as shown in Fig. 1. Inspired by the deep supervision network and feature pyramid network, we add auxiliary segmentation heads at different resolution levels of the decoder to produce the predictions at different scales. To introduce more perturbations in the network, we add the dropout layer before these auxiliary segmentation heads. For an image x, the network $f_\phi(x)$ produces a set of multi-scale predictions $\left[p'_1, p'_2, p'_s, \ldots, p'_{S-1}, p'_S\right]$, where the p'_s is the prediction at s-th scale. Note that a smaller s means higher resolution, and we use S to represent the total number of scales. Then, we rescale these multi-scale predictions to the input size and they are denoted as $[p_1, p_2, p_s, \ldots, p_{S-1}, p_S]$. For the labeled data, the loss function can be formulated as:

$$\mathcal{L}_{sup} = \sum_{s=1}^{S} \alpha_s \mathcal{L}(p_s, y) \tag{1}$$

where $\mathcal{L}(.)$ is a standard supervised learning loss function, such as Dice loss (Milletari et al., 2016), cross entropy loss (Çiçek et al., 2016) or combination loss (Yu et al., 2019). α_s is the weighting factor for scale s.

For the unlabeled data, we introduce a consistency regularization by encouraging the multi-scale predictions to be similar. In detail, we design a pyramid consistency loss to minimize the discrepancy (i.e., variance) among the predictions at different scales. To simplify the calculation, we encourage predictions at different scales to be similar to their average prediction, which not only can reduce the computational complexity from $S(S-1)/2$ to S, but also can be seen as a scale-aware pseudo label. We denote the average

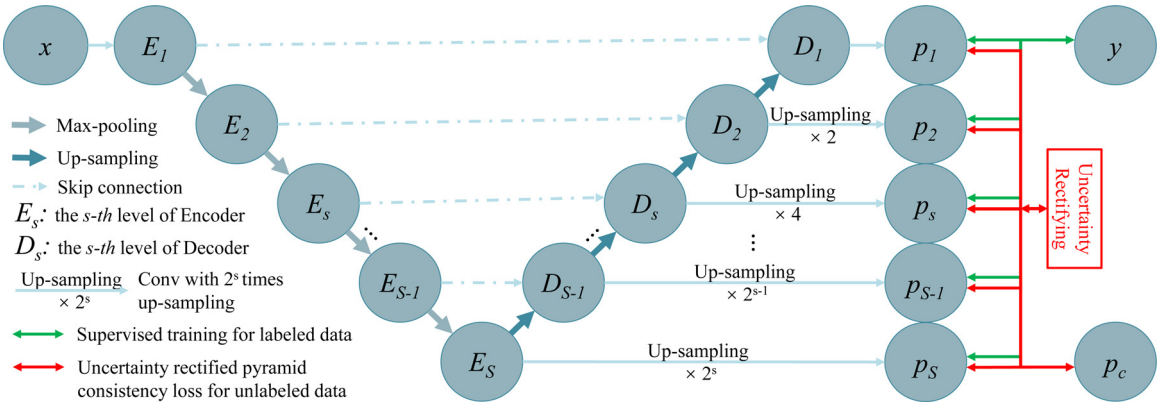

Fig. 1. Overview of the proposed semi-supervised framework via Uncertainty Rectified Pyramid Consistency, which includes a pyramid prediction network and uncertainty estimation module for pyramid consistency rectifying. This framework performs semi-supervised learning by minimizing the supervised loss and the pyramid consistency loss on labeled and unlabeled data, respectively. Considering that there is a lack of labeled data and the up-sampling processing may lead to inaccurate predictions, encouraging them to be consistent at each pixel may be affected by outliers and lead to a performance drop..

prediction as:

$$p_{avg} = \frac{1}{S}\sum_{s=1}^{S} p_s \tag{2}$$

Then, the pyramid consistency loss is formulated as :

$$\mathcal{L}_{pyc} = \frac{1}{S}\sum_{s=1}^{S} \|p_s - p_{avg}\|_2 \tag{3}$$

where we encourage a minimized L_2 distance between the prediction at each scale and the average prediction.

3.2. Uncertainty rectified pyramid consistency

As the pyramid prediction at a range of scales has different spatial resolutions, even they can be re-sampled to the same resolution as the input, the re-sampled results still have different spatial frequencies, i.e., the prediction at the lowest resolution captures the low-frequency component of the segmentation and the prediction at the highest resolution obtains more high-frequency components. Directly imposing a voxel-level consistency among these predictions can be problematic due to the different frequencies, such as loss of fine details or model collapse. Inspired by existing works (Yu et al., 2019; Cao et al., 2020; Wang et al., 2020a; 2019; Zheng and Yang, 2021), we introduce an uncertainty-aware method to address these problems, where a novel uncertainty estimation method based on multi-scale discrepancy is proposed. Differently from the Monte Carlo Dropout-based methods (Yu et al., 2019; Gal and Ghahramani, 2016), our uncertainty estimation leverages the difference between predictions at different scales and it only requires a single forward pass, which needs less computational cost and running time than exiting methods.

3.2.1. Uncertainty estimation based on multi-Scale discrepancy

As the pyramid prediction network can produce multiple predictions at different scales in a single forward pass, uncertainty estimation can be obtained efficiently by measuring their discrepancy without extra effort. To be specific, we use the *KL*-divergence between the average prediction and the prediction at scale s as the uncertainty measurement:

$$\mathcal{D}_s^i \approx \sum_{j=0}^{C-1} p_s^{i,j} \cdot \log \frac{p_s^{i,j}}{p_c^{i,j}} \tag{4}$$

where $\mathcal{D}_s^i$ is the uncertainty of pixel/voxel i at scale s. $p_s^{i,j}$ means the probability of pixel/voxel i belonging to class j in p_s and C is the class number for the segmentation task. The approximated uncertainty shows the pixel/voxel-level difference between the p_s and p_{avg}. Note that for a given pixel/voxel i in $\mathcal{D}_s$, a larger value $\mathcal{D}_s^i$ indicates the prediction for that pixel at scale s is far from the other scales, i.e., with high uncertainty. As result, we obtain a set of uncertainty maps $[\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_s, \ldots, \mathcal{D}_{S-1}, \mathcal{D}_S]$, where $\mathcal{D}_s$ corresponds to uncertainty of p_s.

3.2.2. Uncertainty rectifying

Based on the estimated uncertainty maps $[\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_s, \ldots, \mathcal{D}_{S-1}, \mathcal{D}_S]$, we further extend the pyramid consistency $\mathcal{L}_{pyc}$ to emphasize reliable parts and ignore unreliable parts of the predictions for stable unsupervised training. Specifically, for unlabeled data, we use the estimated uncertainty to automatically select reliable voxels for loss calculation. The rectified pyramid consistency loss is formulated as:

$$\mathcal{L}_{urc} = \frac{1}{S}\sum_{s=1}^{S} \frac{\sum_i ||p_s^i - p_{avg}^i||_2 \cdot w_s^i}{\sum_i w_s^i} \tag{5}$$

where p_s^i and w_s^i are the corresponding prediction and rectifying values for pixel/voxel i at the s-th scale prediction. Follow the policy in (Zheng and Yang, 2021), we define the pixel/voxel-level rectifying values as:

$$w_s^i = e^{-\mathcal{D}_s^i} \tag{6}$$

According to this definition, for a given pixel/voxel at the scale s, a higher uncertainty automatically leads to a lower weight. Differently from many threshold-based cut-off approaches (Yu et al., 2019; Cao et al., 2020), this strategy does not require additional manual efforts to design or tune the threshold carefully. In addition, inspired by previous works (Grandvalet et al., 2005; Zheng and Yang, 2021) that showed reducing the prediction entropy can boost the model's robustness, we further introduce the uncertainty minimization term as a constraint directly. The uncertainty minimization constraint is defined as:

$$\mathcal{L}_{umc} = \frac{1}{S}\sum_{s=1}^{S} \mathcal{D}_s \tag{7}$$

3.2.3. Entire loss for unlabeled images

Based on the rectified pyramid consistency loss and the uncertainty minimization loss, the entire loss function for unlabeled images is defined as:

$$\mathcal{L}_{unsup} = \beta \cdot \mathcal{L}_{urc} + (1.0 - \beta) \cdot \mathcal{L}_{umc} \tag{8}$$

where β is a weight to balance the impact of the two terms. With a combination of these two loss functions for unlabeled images, the segmentation network can focus on the reliable regions, as it is beneficial to reduce the overall uncertainty of the model and produce more consistent predictions across different scales. Here, we would like to point out that $\mathcal{L}_{pyc}$(3) and $\mathcal{L}_{umc}$(7) are two ways of quantifying the discrepancy between p_s and p_{avg}. In this work, we investigated the performance of using each of them or their combinations and presented a complementary combination $\mathcal{L}_{unsup}$ (8).

3.3. The overall objective function

The proposed URPC framework learns from both labeled data and unlabeled data by minimizing the following combined objective function:

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda \cdot \mathcal{L}_{unsup} \tag{9}$$

where $\mathcal{L}_{sup}$ is a joint cross-entropy loss and dice loss. $\mathcal{L}_{unsup}$ is presented in Eq. 8. λ is a widely-used time-dependent Gaussian warming up function (Tarvainen and Valpola, 2017; Yu et al., 2019) to control the balance between the supervised loss and unsupervised loss, which is defined as $\lambda(t) = w_{\max} \cdot e^{(-5(1-\frac{t}{t_{\max}})^2)}$, where $w_{\max}$ means the final regularization weight, t denotes the current training step and $t_{\max}$ is the maximal training step.

4. Experiments and results

4.1. Dataset

In this study, we evaluated the URPC and compared it with several previous works on one in-house dataset and two public datasets, including nasopharyngeal carcinoma (NPC) segmentation dataset (NPC-MRI); whole brain tumor segmentation dataset (BraTS2019) and pancreas segmentation (Pancreas-NIH), all of them are 3D segmentation tasks.

4.1.1. NPC-MRI

The NPC-MRI dataset (Luo et al., 2021b) consists of 258 patients of NPC before radiotherapy which was collected from a local cancer treatment center. The nasopharynx gross tumor volume (GTVnx) and lymph node gross tumor volume (GTVnd) were annotated by an experienced oncologist with 10 years of clinical experiments and checked by an expert group. The mean resolution of the dataset was 1.23 mm×1.23 mm×1.10 mm and the mean dimension was 176 × 286×245. In this study, we randomly select 180, 20, 58 cases for training, validation and testing respectively. For pre-processing, we normalize each scan to zero mean and unit variance.

4.1.2. BraTS2019

The BraTS2019 (Menze et al., 2014) training set consists of 335 scans and each scan includes four modalities (FLAIR, T1, T1ce and T2) and each sequence with an isotropic $1mm^3$ resolution. In this work, we investigate semi-supervised segmentation of the whole tumors from FLAIR images. These scans are randomly split into 250, 25 and 60 scans for training, validation and testing respectively. For pre-processing, we crop the zero intensity region and then re-scale the intensity of each scan to [0, 1].

4.1.3. Pancreas-NIH

The Pancreas-NIH (Roth et al., 2015) dataset includes 82 abdominal CT images with pancreas annotation. Following many existing works (Xia et al., 2020; Luo et al., 2021a), we chose the CT window range of [-125, 275] HU (Zhou et al., 2019a; Luo et al., 2021a) and re-sample them to an isotropic $1mm^3$ resolution, then crop the images centering at the pancreas region with enlarged margins (25 voxels) and finally re-scale the intensity to [0, 1]. Following existing works (Xia et al., 2020; Luo et al., 2021a; Shi et al., 2021), in this study, 62 cases and 20 cases are used for training and testing, respectively.

4.2. Implementation details and evaluation metrics

In this work, all methods are implemented by PyTorch (Paszke et al., 2019) on a Ubuntu18.04 desktop with an NVIDIA GTX1080TI GPU. The backbone segmentation network is 3D-UNet (Çiçek et al., 2016), and we modify it to produce pyramid predictions by adding a prediction layer after each up-sampling block of the decoder as auxiliary segmentation head, where the head is implemented by $1 \times 1 \times 1$ convolution layer followed by softmax. The dropout rate is set to 0.3. The SGD optimizer (weight decay=$1e^{-}4$, momentum=0.9) with Eq. 9 as the loss function. The poly learning rate strategy was used to adjust the learning rate, where the initial learning rate l_i was multiplied by $(1.0 - \frac{t}{t_{\max}})^{0.9}$ where $l_i = 0.1$ and $t_{\max} = 60k$ for the NPC-MRI dataset and $30k$ for the others. The batch sizes were set to 4 for all the compared methods, where half of them are labeled data and the other half are unlabeled images. The model takes randomly cropped patches as input, the patches size is $112 \times 112 \times 112$ on the NPC-MRI dataset, $96 \times 96 \times 96$ on BraTS2019 and Pancreas-NIH datasets. Random cropping, flipping and rotation were used to enlarge the training set and avoid over-fitting. Following previous works (Yu et al., 2019; Luo et al., 2021b), the $w_{\max}$ was also set to 0.1 in this work. In this work, we set the weighting factor α_s and balance weight β to 1 and 0.5, respectively. In the inference stage, the final segmentation results were obtained by using a sliding window strategy. Three widely-used metrics are used to quantitatively evaluate the segmentation performance, including Dice coefficient (DSC), 95% Hausdorff Distance (HD_{95}) and Average Surface Distance (ASD).

4.3. Ablation study

In this study, we performed comprehensive ablation studies in two datasets (Pancreas-NIH and NPC-MRI) to analyze the contributions of each component and further investigate the impact of multi-scale selection for semi-supervised learning.

Firstly, we investigated the performance of URPC for NPC segmentation on the validation set using 18 labeled images and 162 unlabeled images. Our baseline was a naive 3D-UNet (Çiçek et al., 2016) for fully supervised learning. To investigate the number of scales considered for the pyramid consistency, we compared the performance of training with $\mathcal{L}_{pyc}$ when S changes from 1 to 5. Due to the baseline ($S = 1$) can't utilize the unlabeled data, it just uses 18 labeled images for training. The results in the third section of Table 1 show that increasing S from 2 to 4 leads to the performance improvement gradually, but when S is 5, the segmentation performance becomes worse than 4. This is mainly because the prediction at scale 5 has a low resolution with a loss of boundary details. At the same time, we also investigated the impact of $\mathcal{L}_{umc}$ for different S. The result were presented in the second section of Table 1. It can be found that $\mathcal{L}_{umc}$ can bring performance gain for each scale s compared with the baseline. Compared with just using $\mathcal{L}_{pyc}$, the uncertainty rectification module ($\mathcal{L}_{urc}$) only leads to small benefits in the results (0.68% in the mean DSC term). Morever, the combination of $\mathcal{L}_{urc}$ and $\mathcal{L}_{umc}$ achieved the best performance than just using $\mathcal{L}_{pyc}$, $\mathcal{L}_{urc}$ or $\mathcal{L}_{umc}$.

Furthermore, we compared the performances when $S = 4$ with (w) or without (w/o) using $\mathcal{L}_{pyc}$, where w/o $\mathcal{L}_{pyc}$ denotes the network learns from labeled data without using unlabeled data. It can

Table 1
Ablation study of the proposed URPC framework on the NPC MRI validation set, where 18 labeled and 162 unlabeled images were used for training. $\mathcal{L}_{pyc}$, $\mathcal{L}_{urc}$, $\mathcal{L}_{umc}$ denote the pyramid consistency loss, uncertainty rectified pyramid consistency loss and uncertainty minimization constraint loss, respectively..

Method	GTVnx			GTVnd			Mean		
	DSC (%)	ASD (mm)	HD_{95} (mm)	DSC (%)	ASD (mm)	HD_{95} (mm)	DSC (%)	ASD (mm)	HD_{95} (mm)
S = 1 (w/o $\mathcal{L}_{pyc}$)	72.83±11.82	**2.77±1.92**	8.48±9.39	69.34±14.23	5.27±5.96	22.66±19.88	71.09±6.95	4.02±3.81	15.57±13.48
S = 4 (w/o $\mathcal{L}_{pyc}$)	73.47±9.53	2.88±2.61	7.81±8.97	71.22±12.34	5.03±6.22	17.25±16.93	72.35±7.85	3.96±2.81	12.53±15.48
S = 2 (w $\mathcal{L}_{umc}$)	77.79±9.89	2.87±5.17	6.73±5.45	72.33±12.62	6.65±6.87	21.67±18.36	75.06±11.37	4.76±3.95	14.20±7.95
S = 3 (w $\mathcal{L}_{umc}$)	78.69±8.78	2.41±1.42	5.11±3.31	75.05±10.67	4.26±5.25	18.52±14.24	76.87±8.45	3.36±4.08	11.82±9.58
S = 4 (w $\mathcal{L}_{umc}$)	80.18±7.66	1.75±1.24	5.61±4.04	75.33 ±10.98	4.51±5.28	20.41±18.59	77.76±7.57	3.13± 2.12	13.01±12.72
S = 5 (w $\mathcal{L}_{umc}$)	79.74±6.78	1.79±0.84	6.18±4.30	75.34±12.67	4.96±5.25	20.52±19.12	77.54±7.45	3.38±2.84	13.35±14.58
S = 2 (w $\mathcal{L}_{pyc}$)	80.14±6.01	1.48±0.82	5.96±3.63	74.91±12.29	4.03±4.18	23.61±21.82	77.53±8.64	2.75±2.31	14.79±12.45
S = 3 (w $\mathcal{L}_{pyc}$)	79.17±8.51	1.87±1.04	6.83±4.96	76.29±8.54	5.30±5.47	19.73±17.16	77.73±8.46	3.58±2.86	13.28±15.09
S = 4 (w $\mathcal{L}_{pyc}$)	80.93±6.13	1.45±0.57	5.68±3.86	76.79±7.96	3.77±2.51	18.46±17.66	78.86±8.17	2.61±1.87	12.07±11.73
S = 5 (w $\mathcal{L}_{pyc}$)	80.21±7.63	1.59±2.33	6.08±7.52	75.93±11.88	4.95±2.91	21.67±23.59	78.07±7.62	3.27±2.53	13.88±11.53
S = 4 (w $\mathcal{L}_{pyc}$ and $\mathcal{L}_{umc}$)	80.63± 6.78	1.37±2.74	5.99±4.41	77.68±10.35	3.79±5.61	19.84 ±21.72	79.16±8.76	2.58±3.42	12.91±14.67
S = 4 (w $\mathcal{L}_{urc}$)	81.12±5.84	1.25±0.92	5.26±3.58	77.96 ±11.12	4.67±3.61	21.86±23.42	79.54±7.19	2.96±2.38	13.56± 12.91
S = 4 (w $\mathcal{L}_{urc}$ and $\mathcal{L}_{umc}$)	81.35±5.29	1.36±0.89	4.79±3.12	78.48±9.28	4.15±2.87	19.35±17.07	79.91±6.01	2.76±1.69	12.07±9.73

Table 2
Ablation study of the proposed URPC framework on the Pancreas-NIH training set, where 10 labeled and 40 unlabeled images were used for training. $\mathcal{L}_{pyc}$, $\mathcal{L}_{urc}$, $\mathcal{L}_{umc}$ denote the pyramid consistency loss, uncertainty rectified pyramid consistency loss and uncertainty minimization constraint loss, respectively..

Method	DSC (%)	ASD (mm)	HD_{95} (mm)
S = 1 (w/o $\mathcal{L}_{pyc}$)	70.79±22.67	7.47±5.32	21.17±16.92
S = 4 (w/o $\mathcal{L}_{pyc}$)	71.57±17.49	6.28±6.34	20.69±18.26
S = 2 (w $\mathcal{L}_{umc}$)	76.87±9.27	3.16±2.23	8.36±4.83
S = 3 (w $\mathcal{L}_{umc}$)	78.29±8.53	2.97±1.86	8.02±4.34
S = 4 (w $\mathcal{L}_{umc}$)	78.96±6.37	3.19±2.79	9.56±8.77
S = 5 (w $\mathcal{L}_{umc}$)	78.43±7.89	3.27±3.01	11.27±9.73
S = 2 (w $\mathcal{L}_{pyc}$)	77.87±7.65	3.35±1.96	10.27±7.84
S = 3 (w $\mathcal{L}_{pyc}$)	78.86±7.09	3.21±3.42	9.54±11.72
S = 4 (w $\mathcal{L}_{pyc}$)	79.84±5.87	2.68±1.21	8.19±7.28
S = 5 (w $\mathcal{L}_{pyc}$)	79.59±5.95	2.94±1.82	8.07±7.35
S = 4 (w $\mathcal{L}_{pyc}$ and $\mathcal{L}_{umc}$)	80.47±6.19	2.49±2.00	7.86±4.47
S = 4 (w $\mathcal{L}_{urc}$)	80.81±6.58	1.79±1.05	7.59±4.35
S = 4 (w $\mathcal{L}_{urc}$ and $\mathcal{L}_{umc}$)	81.39±5.62	1.99±1.17	6.23±3.49

be found that using $\mathcal{L}_{pyc}$ leads to a large performance improvement from 72.35% to 78.86% in the mean DSC term, as without using $\mathcal{L}_{pyc}$, the pyramid prediction network itself can not leverage unlabeled data for learning. Therefore, we set $S = 4$ in the following experiments. Then, we extended the pyramid consistency to uncertainty rectified pyramid consistency ($\mathcal{L}_{urc}$) and then added the uncertainty minimization term ($\mathcal{L}_{umc}$) to the unsupervised loss, incrementally. To better identify the contribution of the uncertainty rectification module, we also investigated the performance when combining $\mathcal{L}_{pyc}$ and $\mathcal{L}_{umc}$ for the network training. The last section of Table 1 shows that uncertainty rectifying strategy and uncertainty minimization can further improve performance. Compared with the baseline, all these variants achieved large gains, which proved that the pyramid consistency can utilize unlabeled data for better results, and our proposed UPRC achieved the highest performance, with an average *DSC* of 79.91%.

Then, we also investigated the impact of each component and the best value of S on the Pancreas-NIH dataset. As previous methods (Xia et al., 2020; Luo et al., 2021a; Shi et al., 2021) do not have a validation set, we randomly selected 12 cases from the training set as our validation set and used the remaining 50 training cases (10 labeled and 40 unlabeled) for network training. We reported the ablation study results on our selected validation set in Table 2. It shows a similar trend to that in Table 1. We found that the optimal value for S is 4. In addition, the pyramid prediction network with $\mathcal{L}_{pyc}$ achieved better performance than without $\mathcal{L}_{pyc}$ when $S = 4$. Similar to $\mathcal{L}_{pyc}$, $\mathcal{L}_{umc}$ module also improves the baseline's performance by a large margin and achieves the best result in terms of *DSC* when the scale is set to 4. The uncertainty rectifying and uncertainty minimization also brings performance gain. Overall, Table 1 and 2 demonstrated that the pyramid consistency with uncertainty rectifying and uncertainty minimization can improve the baseline performance on different datasets.

4.4. Comparison with existing semi-Supervised methods

We further compared our URPC with five recent semi-supervised learning methods: 1) Mean Teacher (MT) (Tarvainen and Valpola, 2017) that learns by minimizing the difference between the teacher and student predictions of the same input under different perturbations; 2) Interpolation Consistency Training (ICT) (Verma et al., 2019) that is based on mean teacher model and mixup-based data augmentation; 3) Entropy Minimization (EM) (Vu et al., 2019) that utilizes the unlabeled data by reducing the predictions' entropy; 4) Uncertainty Aware Mean Teacher (UAMT) (Yu et al., 2019) that is an extension of mean teacher model with uncertainty estimation and 5) Deep Adversarial Network (DAN) (Zhang et al., 2017) that uses discriminator to classify the labeled and unlabeled data for model regularization. For a fair comparison, all implementations of these methods used the same backbone network are online available. These methods were also compared with simply using the annotated images for supervised learning, which is denoted as SL and serves as a baseline. The experiments were conducted when only 10% and 20% training images were annotated, respectively. The comparison results in different datasets are the following.

4.4.1. Results on NPC-MRI

The quantitative comparisons between our URPC and the other methods on the NPC dataset are presented in Table 3. It is noticeable that all semi-supervised approaches perform better than the supervised learning method (SL) in both cases of 10% and 20% labeled data. DAN (Zheng et al., 2019) and EM (Vu et al., 2019) outperformed other existing methods in 10% and 20% labeled data settings respectively, demonstrating that semi-supervised learning can improve segmentation performance significantly by leveraging unlabeled data. Notably, our URPC achieved better or comparable performance than all existing methods on most evaluation metrics, with significantly higher *DSC* than 5 out of the 6 compared methods. Fig. 2 shows a visual comparison of results obtained by these methods, where DAN (Zheng et al., 2019) and EM (Vu et al., 2019) achieved better results in all existing methods when 10% and 20% training data were annotated, respectively. It demonstrates that our URPC has a higher overlap ratio with the ground truth than EM and DAN and fewer mis-segmentations in both 2D slice-level and

Table 3
Comparison between our method and existing methods on the NPC-MRI dataset. * denotes the p-value < 0.05 based on paired t-test when comparing the proposed with the others.

Labeled %	Method	GTVnx			GTVnd			Mean		
		DSC (%)	ASD (mm)	HD_{95} (mm)	DSC (%)	ASD (mm)	HD_{95} (mm)	DSC (%)	ASD (mm)	HD_{95} (mm)
10%	SL	71.94±11.60*	2.75±2.90*	9.31±9.42*	66.27±14.62*	5.52±10.13*	22.15±24.84*	69.10±10.15*	4.14±5.10*	15.73±13.27*
	MT (Tarvainen and Valpola, 2017)	79.80±6.74*	1.70±1.12	6.35±5.69	69.78±16.33*	6.01±9.70*	22.10±30.59*	74.79±9.15*	3.85±4.83*	14.22±15.86*
	ICT (Verma et al., 2019)	80.58±6.23	1.66±0.92	5.71±4.62	72.62±13.47*	5.01±6.44	20.43±23.81	76.60±7.89*	3.33±3.24	13.07±12.41
	EM (Vu et al., 2019)	79.85±6.32*	1.88±1.55	7.52±10.55*	69.92±12.39*	4.19±7.41	17.77±20.44	74.89±8.85*	3.04±3.71	12.64±11.83
	UAMT (Yu et al., 2019)	79.62±7.16*	1.74±1.03	6.05±4.60	71.98±15.66*	6.09±8.46*	22.82±25.14*	75.80±9.67*	3.91±4.25*	14.44±13.23*
	DAN (Zhang et al., 2017)	80.47±5.73	2.06±2.15*	5.90±4.55	74.62±12.83	3.97±6.34	18.45±21.24	77.54±7.39	3.01±3.27	12.18±10.75
	Ours	80.76±5.72	1.42±0.57	5.79±4.42	75.95±12.74	5.15±8.01	19.88±25.56	78.36±7.66	3.29±4.00	12.83±13.47
20%	SL	80.72±8.28*	1.67±1.65	5.36±3.70	74.16±15.46*	5.17±9.25*	19.32±23.04*	77.44±9.99*	3.42±4.65*	12.34±11.81*
	MT (Tarvainen and Valpola, 2017)	81.10±5.90*	1.64±1.23	5.17±3.06	76.30±13.25*	4.37±7.27*	17.82±22.44	78.70±8.03*	3.00±3.64	11.50±11.35
	ICT (Verma et al., 2019)	81.86±5.91*	1.51±0.74	5.08±3.55	77.42±12.48*	4.97±8.31*	19.64±25.34*	79.64±7.43*	3.24±4.14	12.36±12.90*
	EM (Vu et al., 2019)	82.05±5.28	1.41±0.60	4.95±3.40	77.78±12.18*	4.17±7.72	14.81±20.45	79.92±7.12*	2.79±3.86	9.88±10.34
	UAMT (Yu et al., 2019)	81.38±6.38*	1.41±0.67	5.31±4.02	77.47±12.55*	4.32±7.08*	16.39±20.72	79.43±7.86*	2.87±3.55	10.85±11.08
	DAN (Zhang et al., 2017)	81.68±5.68*	1.60±1.12	5.23±3.45	78.09±12.91*	3.61±6.37	15.09±20.52	79.88±7.41*	2.60±3.18	10.16±10.25
	Ours	82.39±5.67	1.58±1.86	5.76±7.80	79.79±10.81	3.27±5.61	16.37±21.50	81.22±6.43	2.42±2.94	11.07±11.43
100%	Full-Sup	83.93±4.77	1.26±0.54	4.30±3.01	83.10±9.05	2.43±5.22	8.61±12.80	83.51±5.35	1.85±2.61	6.45±6.50

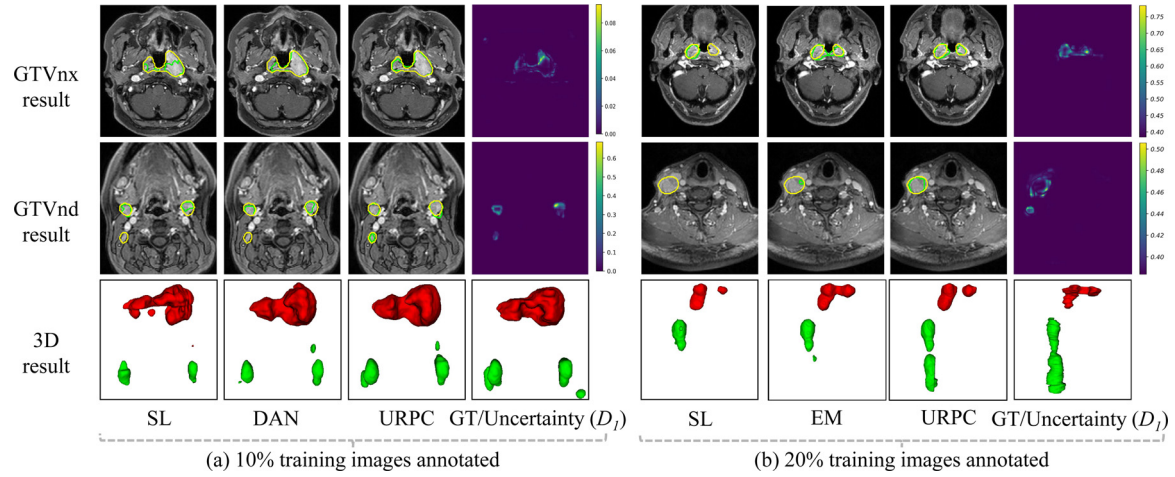

Fig. 2. Visualization of results by different methods and uncertainty map obtained by our method on the MRI-NPC dataset. Lime and yellow contours denote the prediction and ground truth, respectively. In 3D results, the red and green colors show the GTVnx and GTVnd segmentation, respectively.

Table 4
Comparison between our method and existing methods on the BraTS2019 dataset. * denotes the p-value < 0.05 based on paired t-test when comparing the proposed with the others.

Labeled %	Method	DSC (%)	ASD (mm)	HD_{95} (mm)
10%	SL	79.09±15.12*	7.53±10.71*	22.43±26.99*
	MT (Tarvainen and Valpola, 2017)	81.70±14.25*	3.56±4.78*	13.28±15.97*
	ICT (Verma et al., 2019)	82.70±12.33*	4.07±5.78*	13.43±16.53
	EM (Vu et al., 2019)	82.35±13.10*	3.68±4.92*	14.70±17.51*
	UAMT (Yu et al., 2019)	80.93±14.54*	5.43±8.72*	17.71±22.43*
	DAN (Zhang et al., 2017)	82.50±12.44*	3.79±4.58*	15.11±17.70*
	Ours	84.16±11.01	2.63±3.50	11.01±13.37
20%	SL	80.58±14.85*	7.33±10.10*	22.09±26.21*
	MT (Tarvainen and Valpola, 2017)	85.03±11.70	1.89±2.08	7.80±8.59
	ICT (Verma et al., 2019)	84.67±11.97	2.39±3.60*	8.97±11.53
	EM (Vu et al., 2019)	84.82±10.55	3.21±3.96*	12.37±17.20*
	UAMT (Yu et al., 2019)	85.05±11.39	3.03±3.87*	12.31±17.32*
	DAN (Zhang et al., 2017)	84.63±12.79	2.34±2.95	8.96±11.24
	Ours	85.49±10.89	2.04±3.32	8.47±11.41
100%	Full-Sup	88.51±6.90	1.81±2.86	7.52±10.50

3D volume-level. The uncertainty map ($\mathcal{D}_1$) also shown that the high uncertain region is mainly distributed near the boundary.

4.4.2. Results on brats2019

We further evaluated URPC on the BraTS2019 dataset. Table 4 lists the results of all methods on the testing dataset. It can be found that URPC significantly outperformsed all the existing methods when 10% training images were annotated, where ICT (Verma et al., 2019) achieved the best performance than the other existing methods with 82.70% of *DSC* but URPC outperforms it with a gain of 1.47% in *DSC*. Meanwhile, the URPC outperformed all the existing methods in a slight margin when 20% training images were annotated. The visual comparison based on 10% labeled data is presented in Fig. 3. Compared with ICT (Verma et al., 2019) and baseline methods, our URPC model achieved the most accurate segmentation results with fewer over-/under-segmentation regions.

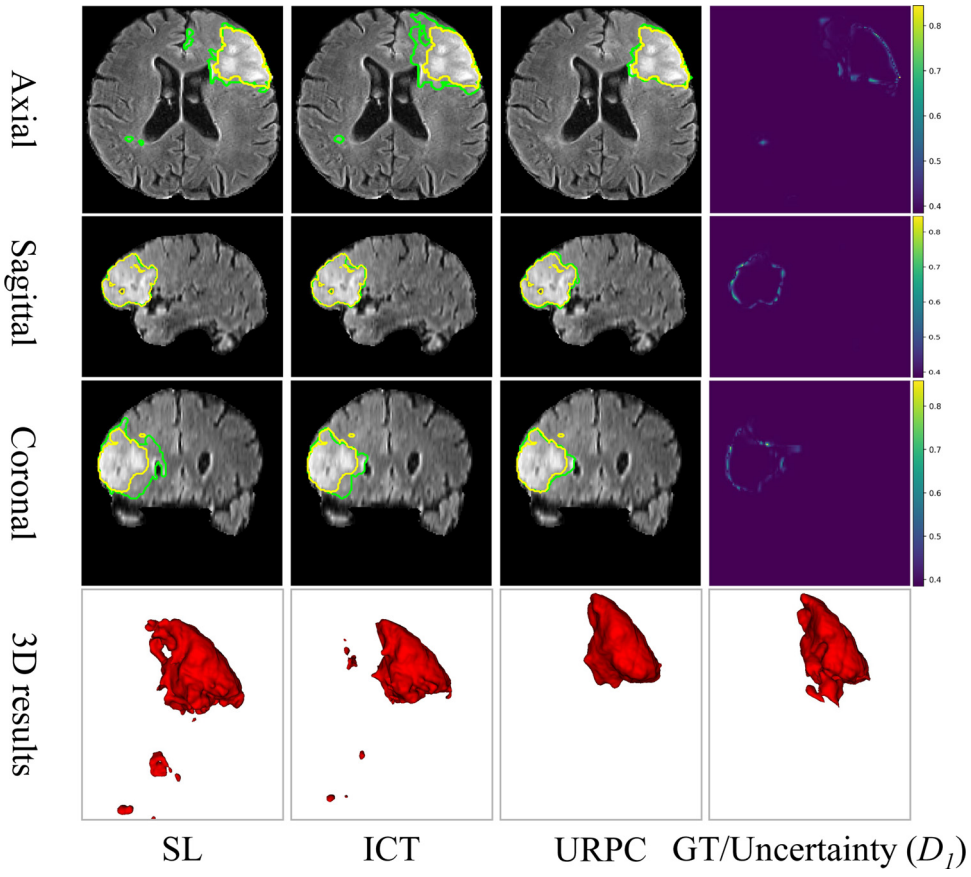

Fig. 3. Visual comparison between different methods for brain tumor segmentation with 10% training data being annotated. Lime and yellow contours denote the prediction and ground truth, respectively.

Table 5
Comparison between our method and existing methods on the Pancreas-NIH dataset. * denotes the *p*-value < 0.05 based on paired *t*-test when comparing the proposed with the others.

Labeled %	Method	DSC (%)	ASD (mm)	HD_{95} (mm)
10%	SL	64.32±13.91*	13.89±6.12*	42.23±15.56*
	MT (Tarvainen and Valpola, 2017)	70.62±14.16*	5.45±3.67*	17.54±17.74
	ICT (Verma et al., 2019)	72.69±12.20*	2.25±1.63	9.25±6.67
	EM (Vu et al., 2019)	70.68±13.22*	5.55±3.67*	21.17±18.90*
	UAMT (Yu et al., 2019)	71.78±13.02*	3.06±2.59	11.75±12.82
	DAN (Zhang et al., 2017)	72.79±13.07*	4.43±2.76	13.96±11.24
	Ours	74.89±10.21	3.74±2.54	11.30±8.07
20%	SL	72.38±20.58*	7.28±4.67*	20.20±17.40*
	MT (Tarvainen and Valpola, 2017)	78.20±8.81*	1.67±0.77	7.41±4.74
	ICT (Verma et al., 2019)	77.37±9.61*	2.13±1.54	8.39±7.17*
	EM (Vu et al., 2019)	76.75±10.39*	3.59±2.68*	12.87±11.50*
	UAMT (Yu et al., 2019)	78.63±8.52	2.91±1.91*	7.81±4.66*
	DAN (Zhang et al., 2017)	79.06±6.58	2.75±2.05	9.61±12.43*
	Ours	80.31±5.77	2.10±1.56	6.58±3.95
100%	Full-Sup	83.01±5.42	1.43±0.60	4.39±1.67

4.4.3. Results on pancreas-NIH

Furthermore, we followed previous works (Xia et al., 2020; Luo et al., 2021a; Shi et al., 2021) to train the network ($S = 4$) with the whole training set (62 cases) for fair comparison with five existing methods. The results are listed in Table 5. As shown in this table, DAN (Zheng et al., 2019) achieved the best performance than all the existing methods in both 10% and 20% labeled ratio settings, but it also is inferior to our proposed URPC. The proposed method outperformed all the compared methods in term of *DSC*, especially in the case of 10% labeled data our method achieved a significant improvement over the second method with a large gain 2.10% of *DSC*. Overall, the URPC achieved the best performance among all the compared methods, which is closer to the upper bound of learning form 100% annotated images. Fig. 4 shows a visual comparison of results generated by the baseline, DAN (Zheng et al., 2019) and the proposed URPC, when 10% training images were labeled. It demonstrates that the proposed URPC is also able to achieve promising segmentation results though the labeled data is limited, where the results obtained by URPC have higher accuracy and fewer false-positive regions.

4.4.4. Computational cost

For the general semi-supervised learning methods, the main difference is about the training strategies, so they may require different time costs for training. In this work, we investigated these methods' computational-cost based on the same soft and hard-

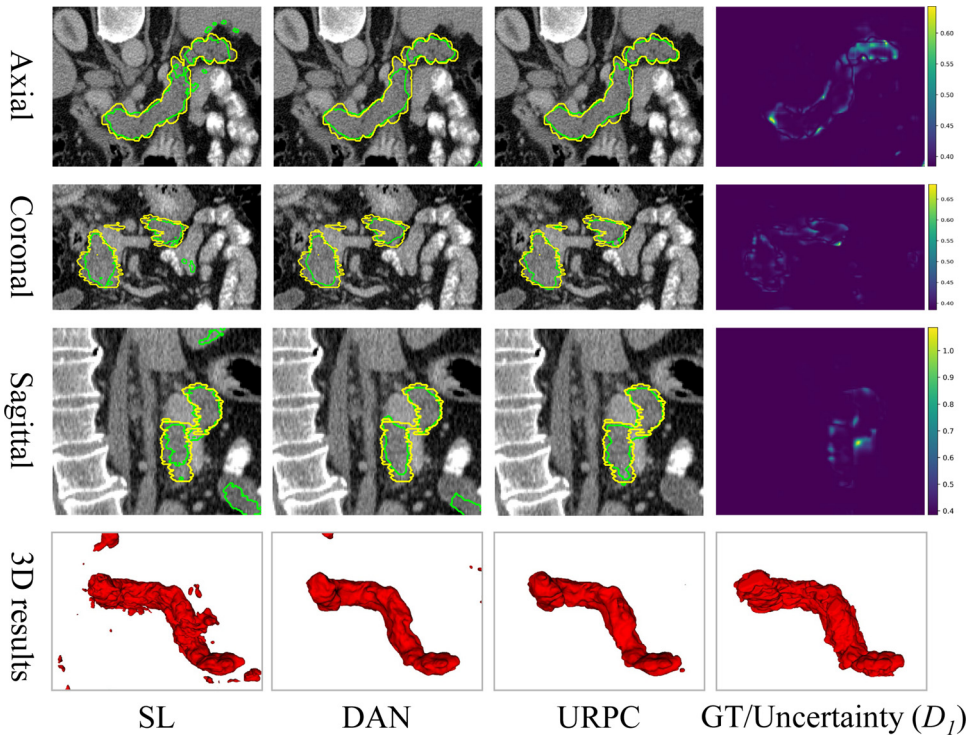

Fig. 4. Visual comparison between different methods for pancreas segmentation with 10% training data being annotated. Lime and yellow contours denote the prediction and ground truth, respectively.

Table 6
Comparison of computational-cost between our method and existing methods based on the MRI-NPC dataset. *FTimes* means the times of an input image passed the networks during one iteration. *TTime* means the total training time (hours).

	SL	MT	ICT	EM	UAMT	DAN	Ours
FTimes	1	2	3	1	9	2	1
TTime (h)	73	76	78	74	95	104	74

ware. We analyzed the number of forwarding pass times of an input image in one iteration and the total training time. It's common sense that with the number of times an input image passes the model increase, the training time and GPU memory also increase rapidly. Table 6 lists the comparison results of computational-cost. It can be found that our framework does not need to perform multiple forward passes or iterative training strategies during the training stage, where most of the existing methods need to pass an image more than two times in an iteration, so our URPC can reduce large computational cost and training time. In addition, all methods require very similar inference costs, as all of them use the same backbone and all auxiliary modules are just used in the training stage.

4.5. Impact of the different ratios of the labeled data

To investigate the data utilization efficiency of the proposed URPC, we performed studies under different ratios of labeled images on the three different datasets and compare the proposed URPC with a fully supervised baseline and a state-of-the-art approach (DAN). The results on NPC-MRI, Pancreas-NIH and BraTS2019 are presented in Fig. 5 (a), (b) and (c) respectively. It is noticeable that both DAN and URPC outperform SL by a huge margin on three datasets, especially when the labeled ratio was 10% and 20%, indicating that semi-supervised methods can achieve promising results when just limited labeled data is accessible. Meanwhile, the proposed URPC consistently performed better than the SL and DAN (Zheng et al., 2019) in different labeled ratio settings, which demonstrates that our method is able to leverage the unlabeled data and bring performance improvement. Furthermore, it can be observed that when increasing the labeled ratio to 50%, our URPC achieves very close results to learning from 100% annotated images. These results indicate that our URPC has the potential to achieve accurate segmentation results with only a small set of training images being labeled, which is desirable for reducing the annotation cost in clinical practice.

4.6. Analysis of hyper-parameter β and consistency regularization

4.6.1. Sensitivity analysis of β

In this part, we first investigated the sensitivity of β in Eq. 5. The β plays a vital role in the proposed method and controls the usage of the uncertainty rectified pyramid consistency term and the uncertainty minimization constraint term. Following the experimental setting of Section 4.3, we investigated the segmentation performance of URPC when the β is set to different values on the NPC-MRI dataset with 18 labeled images and 162 unlabeled images and also on the Pancreas-NIH dataset using 20% labeled data and 80% unlabeled data, respectively. Fig. 6 shows the segmentation performance when the β was set to $\{0.0, 0.1, 0.2, 0.3, 0.5, 0.7, 0.9, 1.0\}$. It can observe that the performance of URPC is not sensitive to the value of β when it is around 0.5.

4.6.2. Analysis of consistency regularization and uncertainty measurement

In general, there are two widely-used methods to quantify the discrepancy or uncertainty between two predictions, named *KL*-divergence or L_2. Firstly, we analyzed the difference between using the L_2 and *KL*-divergence to measure the discrepancy of two probabilities by simulating the segmentation procedure. Let's consider a

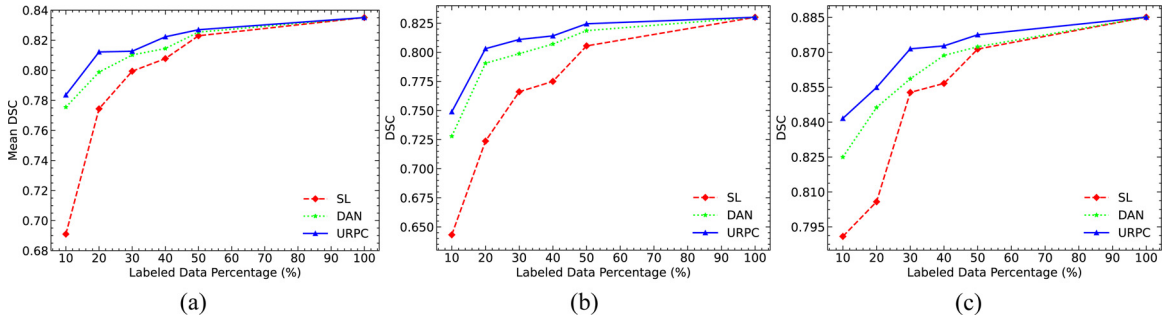

Fig. 5. Performance of three different approaches with different ratios of annotated images. (a) Results of GTVnx and GTVnd segmentation on MRI-NPC. (b) Results of pancreas segmentation on Pancreas-NIH. (c) Results of whole tumor segmentation on BraTS2019.

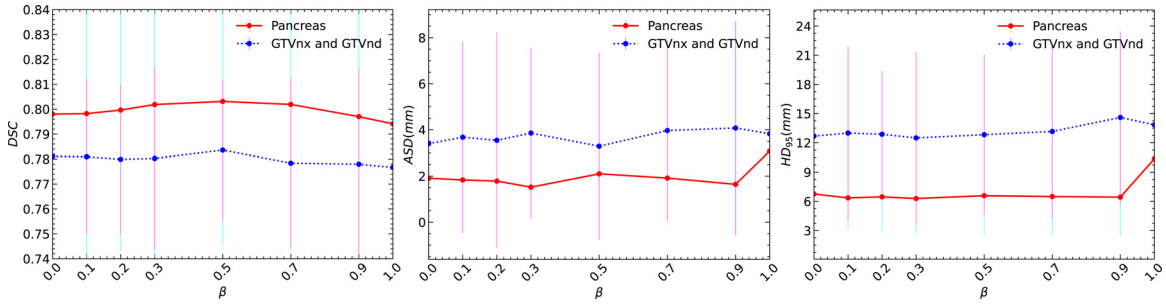

Fig. 6. Sensitivity analysis of hyper-parameter β on the NPC-MRI and Pancreas-NIH datasets..

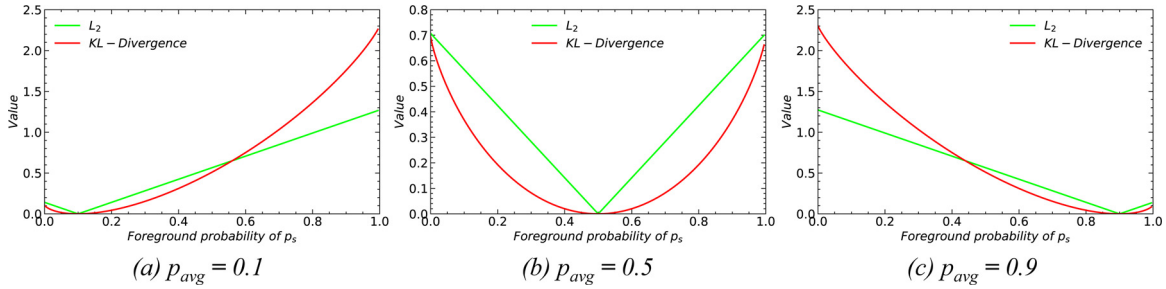

Fig. 7. Toy examples of using L_2 and KL-divergence to measure the discrepancy at a pixel. p_{avg} and p_s represent the foreground probabilities of the average prediction and the prediction at scale s, respectively..

binary segmentation problem at a pixel. Assume the average foreground probability is p_{avg} and p_s is the foreground probability at scale s. We compared the L_2 and KL-divergence-based discrepancy when p_s changes from 0 to 1.0. Fig. 7 shows three toy examples where p_{avg} is set to 0.1, 0.5, 0.9. These toy examples show that the KL-divergence will impose a higher penalty to outliers than L_2. So, the KL-divergence is more discriminative and suitable for the rectification function.

Afterwards, we further investigated the performance of URPC when using KL-divergence or L_2 or their combinations to calculate the consistency regularization loss functions and uncertainty estimation. Following the experimental settings of Section 4.3 and 4.6.1, we conducted the experiments on NPC-MRI and Pancreas-NIH. The quantitative results of using different loss functions combinations on the NPC-MRI and Pancreas-NIH datasets are presented in Table 7. It shows that just using L_2 or KL achieves a worse performance compared with using a joint combination of L_2 and KL.

Furthermore, we also found that using KL to measure the uncertainty and employing L_2 to impose the consistency constraint leads to very similar results compared with utilizing L_2 for uncertainty estimation and KL for consistency regularization. Fig. 8 presents visual examples of using KL and L_2 to estimate the uncertainty map ($\mathcal{D}_1$). It can observe that using KL or L_2 for uncertainty estimation leads to some differences in the boundary or ambiguous regions. It can observe that using KL or L_2 for uncertainty estimation leads to some differences in the boundary or ambiguous regions. It further demonstrates the difference of KL or L_2 in the discrepancy measurement, like Fig. 7 shows. We further presented an example visualization of pancreas segmentation at each scale s, and their corresponding loss function values of $\mathcal{L}_{pyc}$, $\mathcal{L}_{urc}$, $\mathcal{L}_{umc}$ (in Fig. 9). It can be found that the value $\mathcal{L}_{urc}$ is lower than $\mathcal{L}_{pyc}$ at each scale, as the uncertainty rectification module assigns low weights for uncertain pixels. These results show the effectiveness and robustness of the proposed URPC for semi-supervised medical image segmentation.

5. Discussion and conclusion

Despite the deep-learning-based automatic medical image segmentation having achieved great success, it is also limited by requiring a large number of fine annotations when developing clinical applications or tools. Semi-supervised learning learns from limited labeled data and large unlabeled data have shown the potential to deal with this challenge. In this work, we proposed an uncertainty rectified pyramid consistency (URPC) for semi-supervised medical image segmentation. In contrast with existing semi-supervised methods for medical image segmentation (Yu et al., 2019; Bai et al., 2017; Zheng et al., 2019), our URPC extends

Table 7
Analysis of consistency regularization loss functions and uncertainty measurement methods on the NPC-MRI and Pancreas-NIH datasets..

Dataset	$\mathcal{D}_s$ (Eq. 4)	$\mathcal{L}_{urc}$ (Eq. 5)	DSC (%)	ASD (mm)	HD_{95} (mm)
NPC-MRI	L_2	L_2	76.85±3.30	4.52±3.98	13.37±8.78
	KL	KL	77.89±4.18	4.30±3.73	15.53±12.41
	L_2	KL	78.39±3.45	3.84±4.05	14.01±11.5
	KL	L_2	78.36±7.66	3.29±4.00	12.83±13.47
	L_2	L_2	79.62±6.50	2.73±1.59	6.03±3.32
Pancreas-NIH	KL	KL	79.57±6.38	2.83±1.95	7.36±3.75
	L_2	KL	80.20±5.53	2.61±1.82	6.72±3.81
	KL	L_2	80.31±5.77	2.10±1.56	6.58±3.95

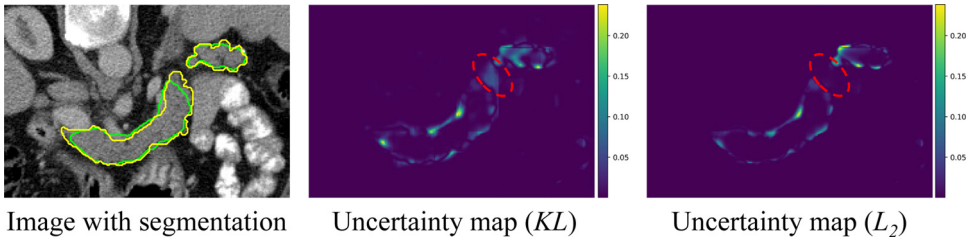

Fig. 8. Visual comparison of estimated uncertainty maps using KL and L_2..

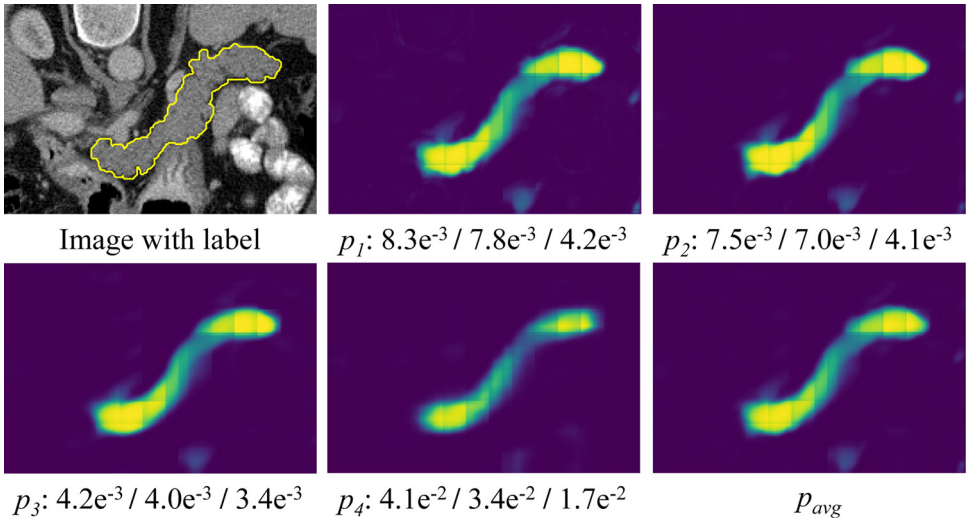

Fig. 9. A example visualization of p_s, p_{avg} and corresponding $\mathcal{L}_{pyc}$ / $\mathcal{L}_{urc}$ / $\mathcal{L}_{umc}$ for each scale s..

the general segmentation network to the pyramid predictions network for semi-supervised image segmentation directly without any complex modules or specific designs. One advantage of the URPC is that it is easy to implement and just needs very minor modifications in the standard segmentation network. Meanwhile, it could be combined with many existing semi-supervised learning methods, such as mean teacher model (Yu et al., 2019), adversarial training (Zheng et al., 2019) and pseudo labels (Bai et al., 2017). Recent semi-supervised segmentation methods mainly focus on designing consistency regularization by adding the perturbation on the input level (e.g., data augmentation), and feature-level (e.g., dropout). But this work presents a multi-scale consistency regularization at the model output level. The proposed method does not conflict with these previous methods, and it can be seen as a multi-scale aware module to extend these existing methods for better performance. However, this work focuses more on evaluating its effectiveness.

Benefiting from multi-scale predictions, we can measure the uncertainty by calculating the variance of these predictions, which requires a single forward pass. Moreover, URPC optimizes for both measures of consistency: the weighted L2 distance ($\mathcal{L}_{urc}$) as well as a form derived from the KL-divergence ($\mathcal{L}_{umc}$). Figs. 7 and 8 show the difference between L_2 and KL-divergence. Table 7 shows that combining the L_2 and KL-divergence to train networks leads to better performance. It shows the applicability of the combined consistency regularization like widely-used combinations of cross-entropy and dice loss functions. In addition, Table 1 and Table 2 show that the uncertainty rectification module brings small performance gains for pyramid consistency. Moreover, the uncertainty minimization constraint is also sufficient to utilize unlabeled data, but the combination of these modules outperforms each submodule and outperforms several existing methods.

Recently, some works, such as multi-head/decoder network and grouped convolution-based CNNs (Wang et al., 2020a) can also produce multiple predictions and estimate the model uncertainty in a single forward pass. However, these methods are limited by the computational cost, as the grouped convolution-based CNN and multi-head/decoder network increase the model capacity and require more GPU memory. In addition, they are designed to deal with interactive refinement and uncertainty estimation, re-

spectively and lack of evaluation in semi-supervised learning. Despite URPC achieving promising results in several datasets, it is also limited by hard generalizing to cross-domain scenarios, as the general semi-supervised learning assumes all samples belong to a similar distribution (Li et al., 2020b). In the future, combing the pyramid consistency with contrastive learning and adversarial training (Vu et al., 2019) may help to handle this challenge (Chaitanya et al., 2020).

In conclusion, this work presents a simple yet efficient uncertainty rectifying pyramid consistency model for semi-supervised medical image segmentation. We first introduce a pyramid consistency regularization to extend the classical fully supervised pyramid prediction network to semi-supervised learning. To encourage the model to learn from reliable regions, we further introduce an uncertainty rectifying strategy to filter unreliable regions automatically. Experiments on two public datasets and one in-house dataset show that the URPC achieves similar or higher accuracy with less computational cost than many recent works, and further indicate the potential of our proposed method to reduce the labeling efforts in clinical workflow. In addition, we release a semi-supervised medical image segmentation codebase and benchmark, which could promote future research in the medical image computing community. In the future, the proposed method can be combined with other consistency-based methods (Luo et al., 2021a; Kervadec et al., 2019; Chen et al., 2019a) to deal with challenging segmentation tasks.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Xiangde Luo: Conceptualization, Methodology, Software, Writing – original draft, Visualization. **Guotai Wang:** Conceptualization, Methodology, Writing – review & editing, Supervision. **Wenjun Liao:** Data curation, Resources. **Jieneng Chen:** Methodology, Writing – review & editing. **Tao Song:** Methodology, Software, Writing – original draft, Visualization. **Yinan Chen:** Resources, Writing – review & editing. **Shichuan Zhang:** Data curation, Resources. **Dimitris N. Metaxas:** Resources. **Shaoting Zhang:** Methodology, Resources, Supervision.

Acknowledgment

This work was supported by the National Natural Science Foundations of China [81771921, 61901084] funding and key research and development project of Sichuan province, China [no. 2020YFG0084]. We would like to thank M.D. Mengwan Wu and Yuanyuan Shen from the Sichuan Provincial Peoples Hospital for the data annotation and checking.

References

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

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A., 2019. Mixmatch: a holistic approach to semi-supervised learning. NeurIPS 32, 5049–5059.

Cao, X., Chen, H., Li, Y., Peng, Y., Wang, S., Cheng, L., 2020. Uncertainty aware temporal-ensembling model for semi-supervised abus mass segmentation. TMI 40 (1), 431–443.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End–to-end object detection with transformers. In: ECCV. Springer, pp. 213–229.

Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E., 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. NeurIPS 33.

Chapelle, O., Scholkopf, B., Zien, A., 2009. Semi-supervised learning. TNN 20 (3). 542–542

Chen, S., Bortsova, G., Juárez, A.G.-U., van Tulder, G., de Bruijne, M., 2019. Multi-task attention-based semi-supervised learning for medical image segmentation. In: MICCAI. Springer, pp. 457–465.

Chen, X., Luo, X., Wangy, G., Zhengy, Y., 2021. Deep elastica for image segmentation. In: ISBI. IEEE, pp. 706–710.

Chen, X., Williams, B.M., Vallabhaneni, S.R., Czanner, G., Williams, R., Zheng, Y., 2019. Learning active contour models for medical image segmentation. In: CVPR, pp. 11632–11640.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: MICCAI. Springer, pp. 424–432.

Cui, W., Liu, Y., Li, Y., Guo, M., Li, Y., Li, X., Wang, T., Zeng, X., Ye, C., 2019. Semi–supervised brain lesion segmentation with an adapted mean teacher model. In: IPMI. Springer, pp. 554–565.

Dou, Q., Yu, L., Chen, H., Jin, Y., Yang, X., Qin, J., Heng, P.-A., 2017. 3D Deeply supervised network for automated segmentation of volumetric medical images. MedIA 41, 40–54.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: ICML, pp. 1050–1059.

Grandvalet, Y., Bengio, Y., et al., 2005. Semi-supervised learning by entropy minimization. NeurIPS 367, 281–296.

Gu, R., Wang, G., Song, T., Huang, R., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S., 2020. Ca-net: comprehensive attention convolutional neural networks for explainable medical image segmentation. TMI 40 (2), 699–711.

Hang, W., Feng, W., Liang, S., Yu, L., Wang, Q., Choi, K.-S., Qin, J., 2020. Local and global structure-aware entropy regularized mean teacher model for 3D left atrium segmentation. In: MICCAI. Springer, pp. 562–571.

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W., Wu, J., 2020. Unet 3+: A full-scale connected unet for medical image segmentation. In: ICASSP. IEEE, pp. 1055–1059.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods 18 (2), 203–211.

Karimi, D., Salcudean, S.E., 2019. Reducing the hausdorff distance in medical image segmentation with convolutional neural networks. TMI 39 (2), 499–513.

Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., Ayed, I.B., 2021. Boundary loss for highly unbalanced segmentation. MedIA 67, 101851.

Kervadec, H., Dolz, J., Granger, E., Ayed, I.B., 2019. Curriculum semi-supervised segmentation. In: MICCAI. Springer, pp. 568–576.

Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z., 2015. Deeply-supervised nets. In: Artificial intelligence and statistics. PMLR, pp. 562–570.

Lee, D.-H., et al., 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML, Vol. 3, p. 896.

Li, S., Sui, X., Luo, X., Xu, X., Liu, Y., Goh, R.S.M., 2021. Medical image segmentation using squeeze-and-expansion transformers. IJCAI 807–815.

Li, S., Zhang, C., He, X., 2020. Shape-aware semi-supervised 3D semantic segmentation for medical images. In: MICCAI. Springer, pp. 552–561.

Li, X., Yu, L., Chen, H., Fu, C.-W., Xing, L., Heng, P.-A., 2020. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. TNNLS 32 (2), 523–534.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: CVPR, pp. 2117–2125.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: CVPR, pp. 3431–3440.

Luo, X., Chen, J., Song, T., Wang, G., 2021. Semi-supervised medical image segmentation through dual-task consistency. AAAI 35 (10), 8801–8809.

Luo, X., Hu, M., Song, T., Wang, G., Zhang, S., 2022. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. MIDL 1–14.

Luo, X., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Chen, N., Wang, G., Zhang, S., 2021. Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In: MICCAI, pp. 318–329.

Luo, X., Wang, G., Song, T., Zhang, J., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S., 2021. Mideepseg: minimally interactive segmentation of unseen objects from medical images using deep learning. MedIA 72, 102102.

Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X., Martel, A.L., 2021. Loss odyssey in medical image segmentation. MedIA 71, 102035.

Masood, S., Sharif, M., Masood, A., Yasmin, M., Raza, M., 2015. A survey on medical image segmentation. Curr. Med. Imaging Rev. 11 (1), 3–14.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al., 2014. The multimodal brain tumor image segmentation benchmark (brats). TMI 34 (10), 1993–2024.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3DV. IEEE, pp. 565–571.

Nie, D., Gao, Y., Wang, L., Shen, D., 2018. Asdnet: Attention based semi-supervised deep networks for medical image segmentation. In: MICCAI. Springer, pp. 370–378.

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, pp. 8026–8037.

Qiao, S., Shen, W., Zhang, Z., Wang, B., Yuille, A., 2018. Deep co-training for semi–supervised image recognition. In: ECCV, pp. 135–152.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: MICCAI, pp. 234–241.
Roth, H.R., Lu, L., Farag, A., Shin, H.-C., Liu, J., Turkbey, E.B., Summers, R.M., 2015. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In: MICCAI. Springer, pp. 556–564.
Roy, A.G., Navab, N., Wachinger, C., 2018. Recalibrating fully convolutional networks with spatial and channel æsqueeze and excitationg blocks. TMI 38 (2), 540–549.
Samuli, L., Timo, A., 2017. Temporal ensembling for semi-supervised learning. In: ICLR, Vol. 4, p. 6.
Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D., 2019. Attention gated networks: learning to leverage salient regions in medical images. MedIA 53, 197–207.
Shi, Y., Zhang, J., Ling, T., Lu, J., Zheng, Y., Yu, Q., Qi, L., Gao, Y., 2021. Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation. IEEE Trans. Med. Imaging 41 (3), 608–620.
Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.-L., 2020. Fixmatch: simplifying semi-supervised learning with consistency and confidence. NeurIPS 33, 596–608.
Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NeurIPS, pp. 1195–1204.
Valvano, G., Leo, A., Tsaftaris, S.A., 2021. Learning to segment from scribbles using multi-scale adversarial attention gates. IEEE Transactions on Medical Imaging 40 (8), 1990–2001.
Verma, V., Lamb, A., Kannala, J., Bengio, Y., Lopez-Paz, D., 2019. Interpolation consistency training for semi-supervised learning. In: IJCAI, pp. 3635–3641.
Vu, T.-H., Jain, H., Bucher, M., Cord, M., Pérez, P., 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: CVPR, pp. 2517–2526.
Wang, G., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S., 2020. Uncertainty-guided efficient interactive refinement of fetal brain segmentation from stacks of MRI slices. In: MICCAI. Springer, pp. 279–288.
Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., 2019. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing 338, 34–45.
Wang, G., Zhai, S., Lasio, G., Zhang, B., Yi, B., Chen, S., Macvittie, T.J., Metaxas, D., Zhou, J., Zhang, S., 2021. Semi-supervised segmentation of radiation-induced pulmonary fibrosis from lung ct scans with multi-scale guided dense attention. IEEE Trans Med Imaging 41 (3), 531–542.
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., 2018. Deepigeos: a deep interactive geodesic framework for medical image segmentation. TPAMI 41 (7), 1559–1572.
Wang, Y., Zhang, Y., Tian, J., Zhong, C., Shi, Z., Zhang, Y., He, Z., 2020. Double-uncertainty weighted method for semi-supervised learning. In: MICCAI. Springer, pp. 542–551.
Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. In: ECCV, pp. 3–19.
Xia, Y., Yang, D., Yu, Z., Liu, F., Cai, J., Yu, L., Zhu, Z., Xu, D., Yuille, A., Roth, H., 2020. Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. MedIA 65, 101766.
Xu, J., Li, M., Zhu, Z., 2020. Automatic data augmentation for 3d medical image segmentation. In: MICCAI. Springer, pp. 378–387.
Yu, L., Wang, S., Li, X., Fu, C.-W., Heng, P.-A., 2019. Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: MICCAI. Springer, pp. 605–613.
Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z., 2017. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: MICCAI. Springer, pp. 408–416.
Zheng, H., Lin, L., Hu, H., Zhang, Q., Chen, Q., Iwamoto, Y., Han, X., Chen, Y.-W., Tong, R., Wu, J., 2019. Semi-supervised segmentation of liver using adversarial learning with deep atlas prior. In: MICCAI. Springer, pp. 148–156.
Zheng, Z., Yang, Y., 2021. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. IJCV 129 (4), 1106–1120.
Zhou, Y., Wang, Y., Tang, P., Bai, S., Shen, W., Fishman, E., Yuille, A., 2019. Semi–supervised 3D abdominal multi-organ segmentation via deep multi-planar co–training. In: WACV. IEEE, pp. 121–140.
Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2019. Unet++: redesigning skip connections to exploit multiscale features in image segmentation. TMI 39 (6), 1856–1867.