An Uncertainty-Aware Transformer for MRI Cardiac Semantic Segmentation via Mean Teachers

Ziyang Wang[1](✉), Jian-Qing Zheng[2], and Irina Voiculescu[1]

[1] Department of Computer Science, University of Oxford, Oxford, UK
ziyang.wang@cs.ox.ac.uk

[2] The Kennedy Institute of Rheumatology, University of Oxford, Oxford, UK

Abstract. Deep learning methods have shown promising performance in medical image semantic segmentation. The cost of high-quality annotations, however, is still high and hard to access as clinicians are pressed for time. In this paper, we propose to utilize the power of Vision Transformer (ViT) with a semi-supervised framework for medical image semantic segmentation. The framework consists of a student model and a teacher model, where the student model learns from image feature information and helps teacher model to update parameters. The consistency of the inference of unlabeled data between the student model and teacher model is studied, so the whole framework is set to minimize segmentation supervision loss and consistency semi-supervision loss. To improve the semi-supervised performance, an uncertainty estimation scheme is introduced to enable the student model to learn from only reliable inference data during consistency loss calculation. The approach of filtering inconclusive images via an uncertainty value and the weighted sum of two losses in the training process is further studied. In addition, ViT is selected and properly developed as a backbone for the semi-supervised framework under the concern of long-range dependencies modeling. Our proposed method is tested with a variety of evaluation methods on a public benchmarking MRI dataset. The results of the proposed method demonstrate competitive performance against other state-of-the-art semi-supervised algorithms as well as several segmentation backbones.

Keywords: Semi-supervised learning · Image semantic segmentation · Vision transformer

1 Introduction

Medical image semantic segmentation is an essential computer vision task with a wide range of applications including robotic surgery, clinical diagnosis, and image alignment. The goal of image semantic segmentation is to classify each pixel of an input image as to whether or not it is part of a Region Of Interest

G. Yang et al. (Eds.): MIUA 2022, LNCS 13413, pp. 494–507, 2022.
https://doi.org/10.1007/978-3-031-12053-4_37

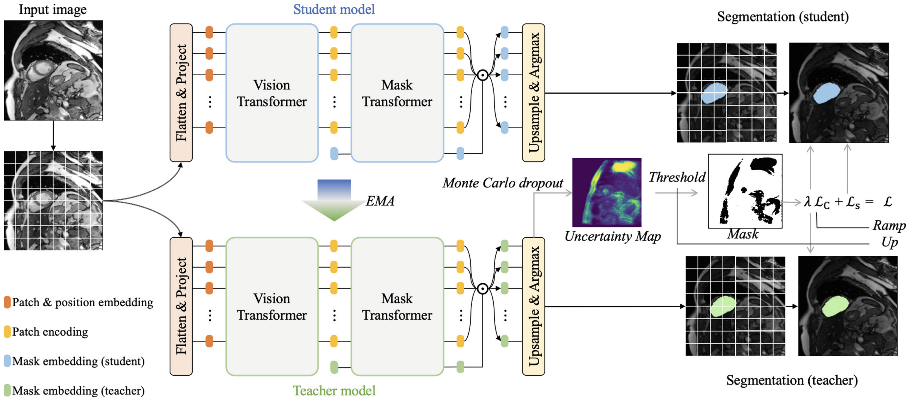

Fig. 1. The framework of semi-supervised uncertainty-aware mean teacher transformer network for medical image segmentation

(ROI) or background. Various deep-learning-based methods haven been widely studied in medical imaging community. The Encoder-Decoder style of Convolutional Neural Network (CNN) is one of the most commonly used segmentation techniques i.e. U-Net [14], and many researchers have studied 3D convolution, atrous convolution, residual learning, attention mechanism with U-Net for a wide range of medical imaging tasks which results in a family of U-Net such as 3D UNet, ResUNet, DenseUNet, Attention-UNet for MRI, ultrasound, CT segmentation [3,5,11,20,21]. There are three main concerns are yet to be further studied: a) the success of deep learning methods relies on a large amount of high-quality annotation data, which is high-cost, time consuming, and difficult to access especially in the clinical domain, b) the semantic feature information cannot be sufficiently condensed and transferred through traditional deep CNN layers or down/up-sampling operations, c) the limitation of the receptive fields in CNNs is not able to model long-range feature information. On order to tacke this challenge, Transformers [18] which use a pure self-attention architecture to model long-range dependencies in natural language processing without CNN are currently studied in the computer vision community. In a similar vein, we propose a ViT network in a semi-supervised manner with uncertainty estimation scheme for medical image semantic segmentation.

We first present a semi-supervised framework that effectively leverages the unlabeled data by encouraging consistent predictions of the same input under different perturbations. Following the Mean Teacher [17] to overcome limitation of Temporal Ensembling [7], the framework consists of the student model and the teacher model where the student model is able to update parameters with gradient descent, and teacher model is updated as an exponential moving average of the student weights. The whole training process is to minimize the segmentation supervision loss between student's machine segmentation (MS) and ground truth (GT), and consistency semi-supervision loss between the teacher's MS and the student's MS. Secondly, inspired by uncertainty estimation [6,23], we utilized Monte Carlo Dropout [6] to estimate the uncertainty with cross-entropy, thus enable student-teacher gradually learn from properly filtering reliable and valuable feature information. And then, to tackle the lack of semantic feature information being transferred through the CNN multi-layers and pooling, we introduce a pure self-attention-based ViT [4] as the semantic segmentation backbone. The segmentation performance benefits from a context model from Natural Language Processing [18], which is also helpful in computer vision especially in pixel-level classification tasks [8]. Finally, the evaluation results demonstrate our method's promising performance against other state-of-the-art semi-supervised methods. Ablation studies include proposed ViT against different CNN-based backbones, several approaches of filtering uncertainty map, and the assumption of different ratio of labeled data provided for training are also explored.

2 Methodology

In the task of semi-supervised learning, $\mathbf{L}$, $\mathbf{U}$, $\mathbf{T}$ normally denote labeled training dataset, unlabeled training dataset, and testing set. We denote a batch of labeled data as $(\boldsymbol{X}, \boldsymbol{Y}_{\mathrm{gt}}) \in \mathbf{L}, (\boldsymbol{X}, \boldsymbol{Y}_{\mathrm{gt}}) \in \mathbf{T}$, and a batch of only raw data as $(\boldsymbol{X}) \in \mathbf{U}$ in unlabeled dataset, where $\boldsymbol{X} \in \mathbb{R}^{h \times w}$ representing a 2D image. $\boldsymbol{Y}_{\mathrm{t}}, \boldsymbol{Y}_{\mathrm{s}}$ are the dense map predicted by the teacher ViT $f_{\mathrm{t}} : \boldsymbol{X} \mapsto \boldsymbol{Y}_{\mathrm{t}}$, and student ViT $f_{\mathrm{s}} : \boldsymbol{X} \mapsto \boldsymbol{Y}_{\mathrm{s}}$, respectively. $\mathcal{L}_{\mathrm{s}} : (\boldsymbol{Y}_{\mathrm{s}}, \boldsymbol{Y}_{\mathrm{gt}}) \mapsto \mathbb{R}, \mathcal{L}_{\mathrm{c}} : (\boldsymbol{Y}_{\mathrm{s}}, \boldsymbol{Y}_{\mathrm{t}}) \mapsto \mathbb{R}$ represent supervised segmentation loss, and semi-supervised consistency loss. In general, the training is to update the parameter of student ViT f_{s} aiming to minimize the combined loss $\mathcal{L}$, which is detailed in Eq. 1. Exponential Moving Average (EMA) [17] is utilized to update parameters of teacher ViT f_{t} from student ViT f_{s} in each training iteration. Uncertainty estimation scheme is applied in $\mathcal{L}_{\mathrm{c}}$ that enable f_{t} to properly guide the training of f_{s} with the certain part of inference. The proposed framework is sketched in Fig. 1. Details of the framework including semi-supervised mean teacher with uncertainty estimation scheme, and segmentation ViT, are discussed in Sects. 2.1 and 2.2.

2.1 Semi-supervised Learning Framework

Inspired by temporal ensembling [7], mean teachers [17], and uncertainty-aware self-ensembling [23], we propose a semi-supervised mean teacher framework with uncertainty estimation scheme for medical image semantic segmentation. The framework is designed to effectively leverage the unlabeled data by encouraging consistent predictions from different perturbations. In each training iteration, the student ViT f_{s} is updated with gradient decent to minimize the combined loss $\mathcal{L}_{\mathrm{s}} + \lambda\mathcal{L}_{\mathrm{c}} = \mathcal{L}$, which is detailed in Eq. 1. λ for $\mathcal{L}_{\mathrm{c}}$ is calculated based on consistency ramp-up method, because it can enable both $f_{\mathrm{s}}, f_{\mathrm{t}}$ can properly make a consistency prediction, and also allow whole framework is able to put more focus on unlabeled data [7] during training process. In the end of each training iteration, EMA is utilized to update parameters of f_{t}, and the prediction of f_{t} is more likely to be correct than f_{s} after a series of study [17].

To further improve semi-supervised performance by enabling f_{t} guide f_{s} to learn feature information via semi-supervised consistency loss $\mathcal{L}_{\mathrm{c}}$, i.e. study on the region where with confident and reliable inference should be utilized to calculate for $\mathcal{L}_{\mathrm{c}}$. We hereby propose uncertainty-aware scheme to enable f_{s} is optimized with $\mathcal{L}_{\mathrm{c}}$ only on confident and reliable inference images. Uncertainty estimation of inference of each pixel, and the approach of filtering the certain/uncertain inference are hereby introduced. Uncertainty estimation is mainly based on the Monte Carlo Dropout [6] on f_{t}, where 8 times stochastic forward passes with dropout and input Gaussian noise. In semantic segmentation task, each pixel is classified with the probability $\boldsymbol{p}$ of ROI, and it is calculated as $\boldsymbol{p} = \frac{1}{T}\sum_t \boldsymbol{p}'_t$ as dropout is utilized, where $\boldsymbol{p}'$ is the probability before dropout. The cross-entropy of predictive $\boldsymbol{U}$ is selected as the metric to estimate the uncertainty of targets [23], and it is calculated as $\boldsymbol{U} = -\sum \boldsymbol{p}\log \boldsymbol{p}$. Therefore, only the region of reliable targets provided by f_{t} (including both ROI and background) are filtered by a threshold τ for f_{s} to be trained with consistency semi-supervision loss $\mathcal{L}_{\mathrm{c}}$, which is detailed in Eq. 2. The supervision segmentation loss $\mathcal{L}_{\mathrm{s}}$ is detailed in Eq. 3.

$$\mathcal{L} = \alpha\mathcal{L}_{\mathrm{s}}(f_{\mathrm{s}}(\boldsymbol{X}), \boldsymbol{Y}_{\mathrm{gt}}) + \lambda\mathcal{L}_{\mathrm{c}}(f_{\mathrm{t}}(\boldsymbol{X}), f_{\mathrm{s}}(\boldsymbol{X})) \tag{1}$$

$$\mathcal{L}_{\mathrm{c}}(f_{\mathrm{t}}(\boldsymbol{X}), f_{\mathrm{s}}(\boldsymbol{X})) = \frac{\|\mathcal{I}(\boldsymbol{U} < \tau) \odot (f_{\mathrm{t}}(\boldsymbol{X}) - f_{\mathrm{s}}(\boldsymbol{X}))^2\|_1}{2\|\mathcal{I}(\boldsymbol{U} < \tau)\|_1 + \epsilon} \tag{2}$$

$$\mathcal{L}_s(f_{\mathrm{s}}(\boldsymbol{X}), \boldsymbol{Y}_{\mathrm{gt}}) = \frac{1}{2}(\mathrm{CrossEntropy}(f_{\mathrm{s}}(\boldsymbol{X}), \boldsymbol{Y}_{\mathrm{gt}}) + \mathrm{Dice}(f_{\mathrm{s}}(\boldsymbol{X}), \boldsymbol{Y}_{\mathrm{gt}})) \tag{3}$$

where $\epsilon = 10^{-6}$, τ is the threshold which is modified in each training iteration based on ramp-up approach. In this way, less data will be removed in training process that enable student model to gradually learn from certain to less certain feature information. λ is a factor for $\mathcal{L}_{\mathrm{c}}$ which is also modified in each training iteration which make the whole framework move focus on minimizing the $\mathcal{L}_{\mathrm{s}}$ to $\mathcal{L}_{\mathrm{c}}$ of training process [23].

2.2 Segmentation Transformer

Semantic feature information is essential in semantic segmentation. The image feature, however, is going to be blurred after multiple layers of CNN encoding. In U-Net, copy and crop are utilized between encoder and decoder to make sufficient semantic feature information been transferred through CNN which results in dominant position in segmentation [14]. The boundary of ROI, especially the information of edge response, can be lost after CNN layers and pooling layers which is harmful for performance [25]. In this section, we introduce a pure self-attention-based vision transformer without CNN for semantic segmentation aiming to achieve sufficient global image context modeling. The model is inspired by Transformer [18], Vision Transformer [4], DETR [2], and Segmentor [16]. The setting of ViT encoder and ViT mask decoder are discussed in this section, and the technical hyper-parameters setting details was introduced in Sect. 3.2.

As shown in Fig. 1, a sequence of patches $\boldsymbol{X}' = [x'_1 \cdots x'_N]^\top \in \mathbb{R}^{N \times P^2}$ is processed from an medical image $\boldsymbol{X} \in \mathbb{R}^{h \times w}$, where P is the patch size, and $N = \frac{h \times w}{P^2}$ is the number of patch from each input image. Each patch is then flatten into a 1D vector and been projected with patch embedding $\boldsymbol{X}_0 = [E_1 \cdots E_N]^\top, E_{1 \ldots N} \in \mathbb{R}^{D \times P^2}$. The positional embeddings to collect positional information $pos = [pos_1 \cdots pos_N]^\top \in \mathbb{R}^{N \times D}$ are added, and the final input sequence of tokens for encoder is $\boldsymbol{Z}_0 = \boldsymbol{X}_0 + pos$. The transformer encoder consists of a multi-headed self-attention (MSA) block followed by a point-wise MLP block of two layers. Residual connections and layer normalization (LN) are both applied in each block. The details of MSA and MLP block for feature learning are demonstrated in Eq. 4, 5, where $i \in 1 \cdots L$, and L is the number of layers in encoder. The self-attention mechanism is composed of three point-wise linear layers mapping tokens to intermediate representations: quires $\boldsymbol{Q}$, keys $\boldsymbol{K}$, and values $\boldsymbol{V}$, which is introduced in Eq. 6. In this way, the transformer encoder maps input sequence $\boldsymbol{Z}_0 = [z_{0,1} \cdots z_{0,N}]$ with position to $\boldsymbol{Z}_L = [z_{L,1}, ..., z_{L,N}]$. All these settings are following by [4]. In this way, the much richer sufficient semantic feature information are fully used in the encoder.

$$\boldsymbol{A}_{i-1} = \mathrm{MSA}(\mathrm{LN}(\boldsymbol{Z}_{i-1})) + \boldsymbol{Z}_{i-1} \tag{4}$$

$$\boldsymbol{Z}_i = \mathrm{MLP}(\mathrm{LN}(\boldsymbol{A}_{i-1})) + \boldsymbol{A}_{i-1} \tag{5}$$

where MSA is calculated by:

$$\mathrm{MSA}(\boldsymbol{Z}') = \mathrm{softmax}(\frac{\boldsymbol{QK}}{\sqrt{D}})\boldsymbol{V}, \tag{6}$$

and the $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ are given by:

$$\boldsymbol{Q} = \mathrm{Linear_Q}(\boldsymbol{Z}'), \boldsymbol{K} = \mathrm{Linear_K}(\boldsymbol{Z}'), \boldsymbol{V} = \mathrm{Linear_V}(\boldsymbol{Z}') \tag{7}$$

The sequence of Z_L is then decoded to dense map $\boldsymbol{S} \in \mathbb{R}^{h \times w \times k}$ as segmentation results via a transformer mask decoder, where k is the number of classes. The decoder acts as mapping patch from encoder and unsample to pixel-level probability of dense map [16]. The learnable class embedding *cls* is processed with $\boldsymbol{Z}_L$ in mask decoder same with transformer encoder with M layers. The output patch sequence is then reshaped to a 2D mask and been bilinearly upsampled to the original image size as prediction results. In transformer mask decoder, both class embedding and patch sequence are jointly processed, and semantic segmentation mask is finally inferenced.

3 Experiments

3.1 Datasets

In this experiment, a MRI cardiac segmentation dataset is selected from the automated cardiac diagnosis MICCAI Challenge 2017 [1]. It consists of 100 different patients divided into 5 evenly distributed subgroups including normal, myocardial infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy, and abnormal right ventricle. We use 44,025 232×256 images from 100 patients. All images are resize to 256×256. 20% of images are selected as testing set, and the rest of dataset is for training. The ratio of assumed labeled data/training set is 10% for direct comparison experiment with similarity measures and difference measures against other semi-supervised methods, other segmentation backbones, and ablation studies, 1%, 2%, 3%, 5%, 10%, 15% and 20% for direct comparison with IOU against other semi-supervised methods.

3.2 Training Details

Our code has been developed under Ubuntu 20.04 in Python 3.8.8 using Pytorch 1.10 [12] and CUDA 11.3 using four Nvidia GeForce RTX 3090 GPU with 24 GB memory, and Intel (R) Intel Core i9-10900K. All the baseline algorithms are directly utilized from [10], and the ViT for segmentation purpose is based on [16] from [15] and TIMM library [22]. The runtime averaged around 3.5 h, including the data transfer, model training, inference and evaluation. All semi-supervised methods are trained with same settings, i.e. training for 30,000 iterations then been tested directly, batch size is set to 24, optimizer is SGD, and learning rate is initially set to 0.01, momentum is 0.9, and weight decay is 0.0001. After multi-times experiments, we finally come up with a proper hyper parameters setting for segmentation ViT which achieve the best results with limited computation resources(6 GB in GPU memory costs): The patch size is 16×16, the number of multi-attention heads is 6, the number of layers L of encoder is 12, normalization method is same with Transformer [18], and the number of layers M of decoder is 2.

3.3 Evaluation

Our proposed semi-supervised method is compared with mean teachers [17], deep adversarial network [24], adversarial entropy minimization for domain adaptation [19], uncertainty-aware self-ensembling model [23], and deep co-training [13] as semi-supervised baseline methods with U-Net [14] as backbone. The direct comparison experiments are conducted with a variety of evaluation metrics including similarity measures: Dice, IOU, Accuracy, Precision, Recall/Sensitivity, Specificity, which are the higher the better. We also investigate difference measures: Relative Volume Difference (RVD), Hausdorff Distance (HD), Average Symmetric Surface Distance (ASSD), which are the lower the better.

3.4 Results

Figure 2 illustrates some examples of raw images, and MS against GT where Yellow, Red, Green and Black represent as True Positive, False Positive, False Negative and True Negative pixel, respectively. Example raw images with uncertainty map, and mask of certain image in three different training stages are illustrated in Appendix. The best result was in **Bold**, and quantitative results are

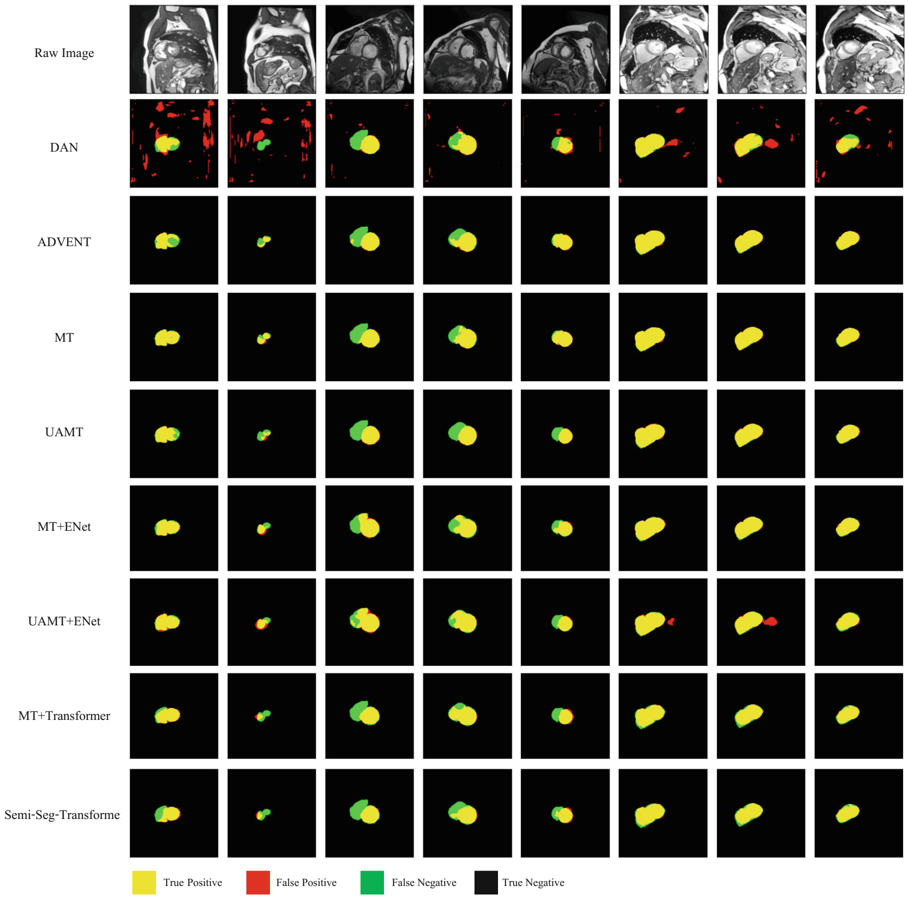

Fig. 2. The example raw images and inference results on testing set (Color figure online)

detailed in Table 1 and Table 2. The evaluation results demonstrate that proposed method promising performance against other semi-supervised methods. Figure 3 gives a systematic review of how the IOU varies when 1%, 2%, 3%, 5%, 10%, 15% and 20% of the training set is labeled. More details of quantitative analysis for different assumed ratio of labeled data given is illustrated in Appendix.

3.5 Ablation Study

In order to analyze the effects of each of the proposed contributions and their combinations, extensive ablation experiments have been conducted. Table 3 annotates with ✓ the use of the mandatory mean teacher for semi-supervise purpose, demonstrating how the removal of uncertainty estimation compromises the overall performance. The model is selected and tested with U-Net [14],

Table 1. Direct comparison with similarity measures on cardiac MRI testing set (the higher, the better)

Model	Dice	IOU	Acc	Pre	Rec/Sen	Spe
[17]	0.8567	0.7494	0.9895	0.7903	0.7903	**0.9977**
[24]	0.5395	0.3694	0.9480	0.4172	0.7631	0.9557
[19]	0.8612	0.7563	0.9896	0.9258	0.8051	0.9973
[23]	0.8347	0.7164	0.9873	0.8683	0.8037	0.9949
[13]	0.8787	0.7836	0.9908	0.9248	0.8370	0.9972
Ours	**0.8821**	**0.7891**	**0.9910**	**0.9288**	**0.8398**	0.9973

Table 2. Direct comparison with difference measures on cardiac MRI testing set (the lower, the better)

Model	RVD	HD	ASSD
[17]	0.3715	28.5797	6.4947
[24]	2.2593	145.4982	49.5673
[19]	0.2669	20.3860	4.7762
[23]	0.3925	27.2209	6.4702
[13]	**0.2630**	21.0363	4.3865
Ours	0.2732	**13.1815**	**3.7085**

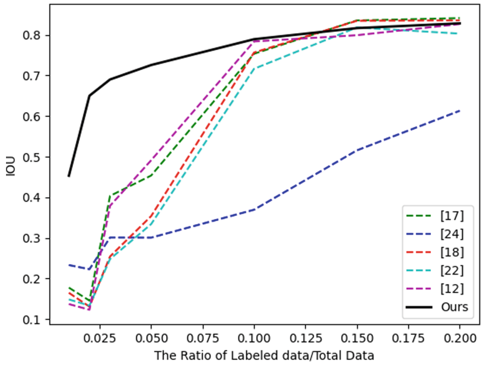

Fig. 3. The IOU performance on test set with different ratio of labeled/total training set

E-Net [12], and proposed segmentation ViT. Further experiments under the assumption of fully supervised learning are also conducted annotated with full ✗ in Table 3. Our proposed ViT with uncertainty estimation scheme shows promising performance especially in IOU and sensitivity in both semi-supervised and fully-supervised manner, respectively. The extended experiments of threshold setting of τ and weight λ of $\mathcal{L}_s$ in training process is illustrated in Appendix.

Table 3. Ablation studies on contributions of architecture and modules (the higher, the better)

Mean teacher	Uncertainty aware	Model	IOU	Sen	Spe
✓		UNet	0.7494	0.7903	**0.9977**
✓	✓	UNet	0.7164	0.8037	0.9949
✓		ENet	0.7549	0.8314	0.9958
✓	✓	ENet	0.7460	0.8529	0.9941
✓		Ours	0.7840	0.8405	0.9970
✓	✓	**Ours**	**0.7891**	**0.8398**	0.9973
✗	✗	UNet	0.7924	0.8409	**0.9975**
✗	✗	ENet	0.7549	0.8696	0.9937
✗	✗	**Ours**	**0.8173**	**0.9137**	0.9951

4 Conclusion

Our semi-supervised uncertainty-aware segmentation is successful in using ViT for medical image semantic segmentation via a mean teacher framework. Experimental results on the public MRI dataset demonstrate our method's promising performance compared against both supervised and semi-supervised existing methods. In the future, multi-task learning and multi-view learning which potentially improve semi-supervised learning performance will be further studied.

A Appendix

Table 4 gives detailed systematic IOU results under different assumptions of the ratio of labeled to total data, on the MRI Cardiac test set. It is pleasantly remarkable to see serviceable results being obtained with a proportion of labelled data as small as 1%, 2%, or 3% of the total. Given the small set of type-specific annotations that exist, they can now be put to good use by pairing them with large amounts of unlabeled data and making them available through our proposed method.

Table 4. The IOU results under different assumption of ratio of label/total data on MRI cardiac test set (the higher, the better)

	1%	2%	3%	5%	10%	15%	20%
[17]	0.1776	0.1457	0.4034	0.4536	0.7533	**0.8354**	**0.8411**
[24]	0.2331	0.2230	0.3010	0.3007	0.3694	0.5155	0.6130
[19]	0.1649	0.1309	0.2543	0.3538	0.7563	0.8345	0.8356
[23]	0.1486	0.1334	0.2480	0.3341	0.7163	0.8180	0.8029
[13]	0.1372	0.1232	0.3790	0.4912	0.7836	0.7990	0.8265
Ours	**0.4531**	**0.6500**	**0.6900**	**0.7256**	**0.7891**	0.8165	0.8282

Table 5 and Table 6 reports the different approaches to modify the threshold τ of filtering certain or uncertain pixels with uncertainty estimation scheme, and the weight λ of loss $\mathcal{L}_c$ in each training iteration. We explore the fixed value, exponential ramp up [7], linear ramp up, cosine ramp down [9] and variants of them. Details of exponential ramp up, linear ramp up and cosine ramp down is illustrated in the following Eq. 8, 9, 10, respectively. Each experiment is conducted with different approaches under the other one either τ or λ is fixed with exponential ramp up. The results illustrates different approaches of updating τ, λ in each training iteration cannot significantly improve the performance of proposed method, and all other experiments for τ, λ is with exponential ramp up.

$$\tau or \lambda = e^{-5\times(1-t_{\text{iteration}}/t_{\text{maxiteration}})^2} \tag{8}$$

$$\tau or \lambda = t_{\text{iteration}}/t_{\text{maxiteration})} \tag{9}$$

$$\tau or \lambda = 0.5 \times (cosine(\pi \times t_{\text{iteration}}/t_{\text{maxiteration}}) + 1) \tag{10}$$

Table 5. Ablation studies on the threshold setting of uncertainty in training process (the higher, the better)

Threshold	Model	IOU	Acc	Pre	Sen	Spe
Threshold 0.2	UNet	0.7465	0.9889	0.8895	0.8229	0.9958
Threshold 0.5	UNet	0.7480	0.9891	0.9048	0.8119	0.9965
Threshold 0.8	UNet	0.7042	0.9862	0.8299	0.8231	0.9930
Exponential Ramp Up	UNet	0.7543	0.9892	0.8895	0.8324	0.9957
Linear Ramp Up	UNet	0.7179	0.9866	0.8189	0.8534	0.9922
Cosine Ramp Down	UNet	0.7046	0.9861	0.8230	0.8305	0.9926
0.6 * Exponential Ramp Up	UNet	0.7321	0.9879	0.8588	0.8324	0.9943
0.6 * Linear Ramp Up	UNet	0.7354	0.9883	0.8852	0.8130	0.9956
0.6 * Cosine Ramp Down	UNet	0.8552	0.9889	0.8931	0.8205	0.9959
0.8 * Exponential Ramp Up	UNet	0.7240	0.9874	0.8528	0.8275	0.9941
0.8 * Linear Ramp Up	UNet	0.7326	0.9882	0.8836	0.8109	0.9956
0.8 * Cosine Ramp Down	UNet	0.7674	0.9899	0.9017	0.8374	0.9962
1.2 * Exponential Ramp Up	UNet	0.7326	0.9882	0.8834	0.8109	0.9956
1.2 * Linear Ramp Up	UNet	0.7304	0.9876	0.8458	0.8426	0.9936
1.2 * Cosine Ramp Down	UNet	0.7493	0.9889	0.8807	0.8340	0.9953
1.4 * Exponential Ramp Up	UNet	0.8359	0.9874	0.8724	0.8024	0.9951
1.4 * Linear Ramp Up	UNet	0.8167	0.9856	0.8305	0.8034	0.9932
1.4 * Cosine Ramp Down	UNet	0.7427	0.9884	0.8638	0.8412	0.9945

Table 6. Ablation studies on the weight setting of consistency loss in training process (the higher, the better)

Weight	Model	IOU	Acc	Pre	Sen	Spe
Threshold 0.2	UNet	0.5243	0.9723	0.6238	0.7667	0.9808
Threshold 0.5	UNet	0.3956	0.9567	0.4719	0.7101	0.9670
Threshold 0.8	UNet	0.4052	0.9703	0.6667	0.5082	0.9894
Exponential Ramp Up	UNet	0.7105	0.9870	0.8613	0.8023	0.9946
Linear Ramp Up	UNet	0.7149	0.9868	0.8357	0.8319	0.9932
Cosine Ramp Down	UNet	0.7547	0.9894	0.9044	0.8201	0.9964
0.6 * Exponential Ramp Up	UNet	0.7723	0.9900	0.8978	0.8467	0.9960
0.6 * Linear Ramp Up	UNet	0.7586	0.9896	0.9069	0.8227	0.9965
0.6 * Cosine Ramp Down	UNet	0.7742	0.9900	0.8908	0.8554	0.9956
0.8 * Exponential Ramp Up	UNet	0.7110	0.9864	0.8216	0.8408	0.9924
0.8 * Linear Ramp Up	UNet	0.7248	0.9875	0.8559	0.8256	0.9942
0.8 * Cosine Ramp Down	UNet	0.7178	0.9869	0.8376	0.8338	0.9933
1.2 * Exponential Ramp Up	UNet	0.7432	0.9887	0.8854	0.8223	0.9956
1.2 * Linear Ramp Up	UNet	0.5596	0.9742	0.6363	0.8227	0.9805
1.2 * Cosine Ramp Down	UNet	0.7509	0.9891	0.8955	0.8230	0.9960
1.4 * Exponential Ramp Up	UNet	0.6968	0.9864	0.8621	0.7482	0.9948
1.4 * Linear Ramp Up	UNet	0.6557	0.9832	0.7807	0.8037	0.9906
1.4 * Cosine Ramp Down	UNet	0.7550	0.9893	0.8979	0.8259	0.9961

Figure 4 sketches randomly selected raw images with their corresponding uncertainty maps, and masks generated by proposed method at three different stages (from the beginning to the end) of the training process. In uncertainty maps, yellow represents the teacher ViT f_{t} is uncertain of prediction with the given pixels, and blue represents the teacher ViT f_{t} is certain of prediction with the given pixels. The uncertainty map is gradually moving from yellow to green in the training process as shown in Fig. 4. The threshold of certainty estimation is then applied with uncertainty map which results in masks, where the white represents that the prediction by teacher ViT f_{t} is certain enough to guide the student ViT f_{s} i.e. for calculation the consistency loss $\mathcal{L}_{\mathrm{s}}$, and the black represents that the pixels with uncertainty is temporally unavailable to be considered in consistency semi-supervision loss calculation. Please remind that both the background and ROI can be certain with the white simultaneously. Some typical example masks illustrates that model is only uncertain with the boundary of ROI as shown in Fig. 4, and finally the framework is very likely to be certain with the whole image with a proper threshold setting, that the uncertainty map is going to be blue, mask is going to be white in the end of training process.

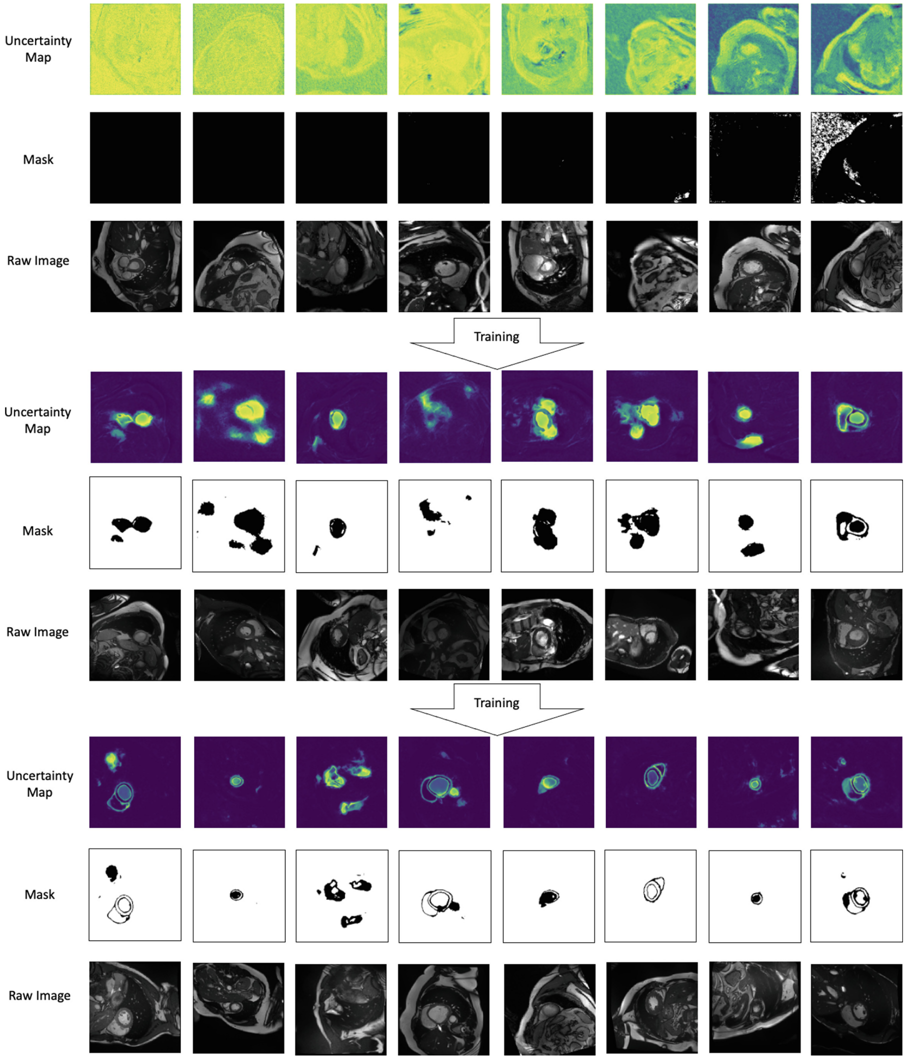

Fig. 4. Sample uncertainty maps, masks, and raw images during the training process (Color figure online)

References

1. Bernard, O., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging **37**(11), 2514–2525 (2018)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13

3. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
4. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Ibtehaz, N., Rahman, M.S.: MultiResUNet: rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural Netw. **121**, 74–87 (2020)
6. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? Adv. Neural. Inf. Process. Syst. **30**, 5574–5584 (2017)
7. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 (2016)
8. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)
9. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
10. Luo, X.: SSL4MIS (2020). https://github.com/HiLab-git/SSL4MIS
11. Oktay, O., et al.: Attention U-Net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
12. Paszke, A., Chaurasia, A., Kim, S., Culurciello, E.: ENet: a deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147 (2016)
13. Qiao, S., Shen, W., Zhang, Z., Wang, B., Yuille, A.: Deep co-training for semi-supervised image recognition. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11219, pp. 142–159. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01267-0_9
14. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
15. Strudel, R.: Segmenter (2021). https://github.com/rstrudel/segmenter
16. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: transformer for semantic segmentation. arXiv preprint arXiv:2105.05633 (2021)
17. Tarvainen, A., Valpola, H.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 1195–1204 (2017)
18. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
19. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: ADVENT: adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2517–2526 (2019)
20. Wang, Z., Voiculescu, I.: Quadruple augmented pyramid network for multi-class COVID-19 segmentation via CT. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC) (2021)
21. Wang, Z., Zhang, Z., Voiculescu, I.: RAR-U-Net: a residual encoder to attention decoder by residual connections framework for spine segmentation under noisy labels. In: 2021 IEEE International Conference on Image Processing (ICIP), pp. 21–25. IEEE (2021)

22. Wightman, R.: Pytorch image models (2019). https://github.com/rwightman/pytorch-image-models. https://doi.org/10.5281/zenodo.4414861
23. Yu, L., Wang, S., Li, X., Fu, C.-W., Heng, P.-A.: Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 605–613. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_67
24. Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z.: Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 408–416. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_47
25. Zhang, Z., Li, S., Wang, Z., Lu, Y.: A novel and efficient tumor detection framework for pancreatic cancer via CT images. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 1160–1164. IEEE (2020)