Inherent Consistent Learning for Accurate Semi-supervised Medical Image Segmentation

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

Semi-supervised medical image segmentation has attracted much attention in recent years because of the high cost of medical image annotations. In this paper, we propose a novel Inherent Consistent Learning (ICL) method to progressively learn the semantic consistency of category representations from labeled and unlabeled data. In practice, we introduce two external modules namely Supervised Semantic Proxy Adaptor (SSPA) and Unsupervised Semantic Consistent Learner (USCL) to realize the above purpose by aligning the category representations of labeled and unlabeled data, as well as updating the global category representations over the entire training set. Once the model is optimized, these external modules are discarded to maintain the simplicity of the main architecture. The proposed ICL is a plug-and-play scheme suitable for various network architectures. Extensive experimental results on three public benchmarks show the proposed method outperforms the state-of-the-art methods by a margin, especially when the number of annotated data is extremely limited. The code is available.

Keywords: Semi-supervised Learning, Medical Image Analysis, Semantic Segmentation

1. Introduction

2. Related Work

2.1. Semi-superivsed Medical Image Segmentation.

Various method and backbone networks have been proposed in recent years towards segmentation tasks, among which the Vision-Transformer-based approaches (Strudel et al., 2021; Chen et al., 2021a; Hatamizadeh et al., 2022) have attracted widespread attention for their strong generalization and great performance. However, these approaches are highly dependent on the number and annotation accuracy of the training data. While non-professionals can annotate general image dataset, the labeling of the medical imaging data must be done by 2˜3 middle- and senior-level professional imaging doctors to ensure the labeling quality (Jiao et al., 2022; Ji et al., 2022).

Recently, many semi-supervised segmentation methods have been proposed to alleviate the dependency on the labeled data of general and medical images. Existing commonly used semi-supervised medical image segmentation methods are mainly related to pseudo-labeling (self-training), co-training, and consistency regularization. Self-training strategy was first proposed by (Blum and Mitchell, 1998) in 1965 and later followed by other variants (Zou et al., 2020; Chen et al., 2021b); its main idea is to find a way to refine and use the generated pseudo labels as data expansion for iterative training. Co-training methods (Blum and Mitchell, 1998; Zhou et al., 2019; Peng et al., 2020; Wang et al., 2021; Luo et al., 2021)

© 2023 Anonymous.

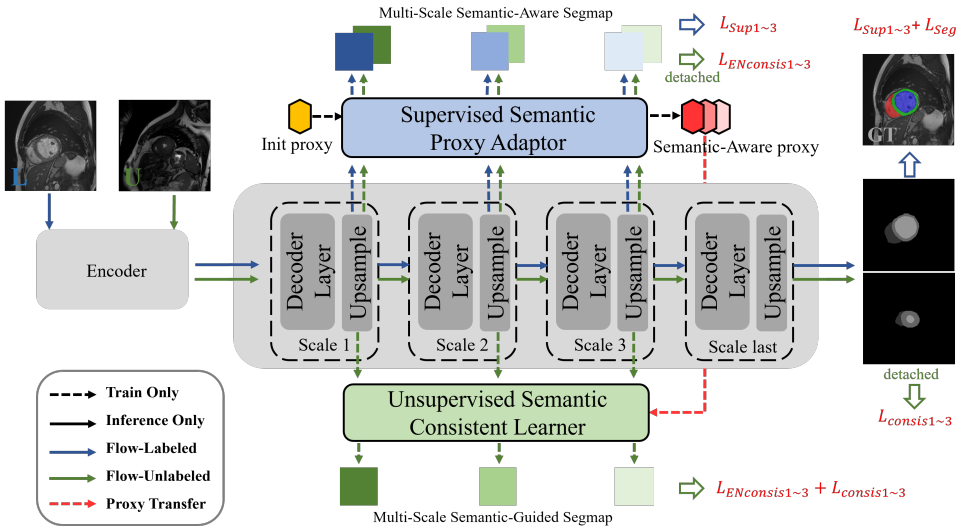

Figure 1: Overview of our proposed Inherent Consistent Learning framework.

assume that different models can learn multiple independent and complementary views from the same dataset. By combining information from multiple views of different perspectives, the accuracy of the final classification can be greatly improved. In (Bachman et al., 2014; Ouali et al., 2020; Luo et al., 2022), consistency regularization was introduced to semi-supervised learning, which they believed that after applying perturbation to an input, its intermediate representations should be consistent. Such strategies can make the model robust to perturbations, generalizing better while learning the inherent invariance of the data.

3. Methodology

For the semi-supervised learning setting, the training set usually contains a small portion of labeled set $D_l = \{(x_i^l, y_i^l)\}_{i=1}^N$ and a much larger unlabeled set $D_u = \{x_i^l\}_{i=1}^M$ (N >> M). The goal is to train a segmentation network by mining information from a small amount of labeled data and efficiently utilizing large amounts of unlabeled data. In this work, the proposed Inherent Consistent Learning (IRL) framework gives a new perspective for semi-supervised segmentation, as illustrated in Fig. 1, which consists of a single encoder-decoder model with a segmentation head f, two newly proposed external modules namely Supervised Semantic Proxy Adaptor (SSPA) and Unsupervised Semantic Consistent Learner (USCL) and a initialized semantic-aware proxy $Q = \mathbb{R}^{Z\times 4C}$, Z indicates the number of classes, aiming at transferring and aligning the category representations of labeled and unlabeled data. During the training phase, both labeled and unlabeled data are fed to the encoder-decoder backbone, generating multi-scale intermediate features f_l and f_u. The labeled features f_l are then sent to the SSPA module with the initialized semantic-aware proxy. Supervised by ground truth, the SSPA can explicitly extract the inherent semantic information through the proxy and the generated multi-scale semantic-aware segmentation maps. By transferring the learned proxy as guidance to the USCL module along with the f_u,

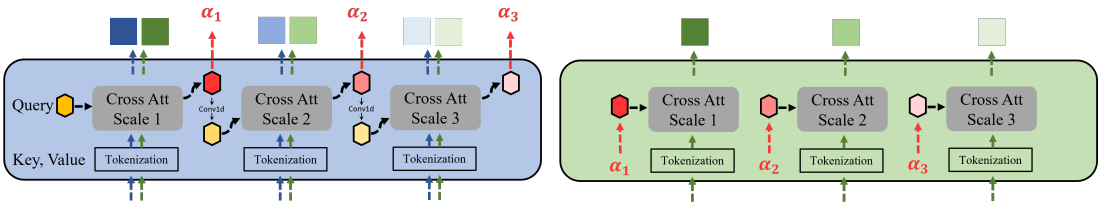

Figure 2: Supervised Semantic Proxy Adaptor and Unsupervised Semantic Consistent Learner modules.

semantic-guided segmentation maps can be generated and later used to perform consistency regularization with the predictions of the unlabeled data. In addition, the unlabeled data can also go through the SSPA module to produce more accurate semantic-aware maps to better form the USCL module with the consistency learning. It is worth noting that such strategy are suitable for both 2D and 3D networks, and both of the external modules are discarded during the inference phase.

3.1. Supervised Semantic Proxy Adaptor

The basic idea of the Supervised Semantic Proxy Adaptor is inspired by TransSeg (Xie et al., 2021). Here we utilize the similar strategy to define a set of learnable semantic-aware proxy as query, and calculate the attention maps with decoder output multi-scale features. With a small conv segmentation head applied to attention maps, final predictions can be obtained and supervised by the ground truth to form the learnable semantic-aware proxy and later use it for guided segmentation with unlabeled data in USCL module.

For simplicity, we use 2D network as an illustration. We first feed a mini-batch of equally size labeled and unlabeled images $\{X_l, X_u\} \in \mathbb{R}^{1\times H\times W}$ to the network. The network decoder outputs a set of multi-scale intermediate features $\{f_l, f_u\}_\lambda, \lambda \in (1, 2, 3)$ from both labeled and unlabeled images. The features are then tokenized and sent to a series Cross-Attention Blocks within SSPA module as input Key and Value $\{t_l, t_u\}_\lambda$, and the the initialized proxy $Q \in \mathbb{R}^{Z\times 4C}$ as the input Query.

As depicted in Fig. 2(left), the initialized proxy is first sent to the cross attention module to conduct interaction with the tokenized features $\{t_l, t_u\} \in \mathbb{R}^{4C\times(\frac{H}{16}\times\frac{W}{16})}$ at the first scale, generating the semantic-aware proxy $Q_1 \in \mathbb{R}^{Z\times 4C}$ and the semantic-aware attention maps $\{A_{l1}, A_{u1}\} \in \mathbb{R}^{Z\times(\frac{H}{16}\times\frac{W}{16})}$. The proxy is then sent to the next scale of cross-attention after being applied a 1×1 conv to reduce the dimension from 4C to 2C, obtaining the semantic-aware attention map at scale 2 and so on. Each scale of attention map $\{A_{l\lambda}, A_{u\lambda}\} \in \mathbb{R}^{Z\times(\frac{\lambda H}{16}\times\frac{\lambda W}{16})}$ can be applied a small conv2d segmentation head and get the final predictions $\{P_{l\lambda}, P_{u\lambda,}\} \in \mathbb{R}^{Z\times\frac{\lambda H}{16}\times\frac{\lambda W}{16}}$. The predictions from the labeled data $P_{l\lambda} \in \mathbb{R}^{Z\times\frac{\lambda H}{16}\times\frac{\lambda W}{16}}$ can be later supervised by ground truth, aiming to have a strong discrimination for the inherent semantic representations for different categories.

In addition, at each scale, the semantic-aware proxy $Q_\lambda \in \mathbb{R}^{Z\times\frac{4C}{\lambda}}$ and the detached semantic-aware segmentation map $P_{u\lambda} \in \mathbb{R}^{Z\times(\frac{\lambda H}{16}\times\frac{\lambda W}{16})}$ will be sent to the USCL module and perform guided segmentation with the the well-learned global category representations

and consistency supervision with the semantic-guided segmentation maps from the USCL module.

3.2. Unsupervised Semantic Consistent Learner

Unlike SSPA module that performs progressively interaction with updated semantic-aware proxy, in this module, each scale of cross attention module receives a corresponding proxy from SSPA as semantic-aware guidance $Q_\lambda \in \mathbb{R}^{Z \times \frac{4C}{\lambda}}$ as mentioned in section **3.1**.

As depicted in Fig. 2(right), these semantic-aware proxies interact with the unlabeled features and generate semantic-guided attention maps $A_{u\lambda} \in \mathbb{R}^{Z \times (\frac{\lambda H}{16} \times \frac{\lambda W}{16})}$ at different scale $\lambda \in (1, 2, 3)$. These multi-scale attention maps are utilized to generate final predictions similar to SSPA module. Such predictions will be later used to perform consistency learning with the final detached output of the main network. We consider this as a new strategy to efficiently leverage the large amounts of unlabeled data. In addition, since the detached semantic-aware segmentation maps are generated through the well-supervised SSPA module, they can provide more reliable consistency regularization to better form the USCL module.

4. Experiments

4.1. Datasets and evaluation metrics

In this work, we perform extensive experiments on three public datasets, including **ACDC** (Bernard et al., 2018), **BraTS** (Menze et al., 2014) and **AMOS** (Ji et al., 2022).

ACDC dataset consists of 200 short-axis cardiac cine-MR images from 100 patients with three segmentation targets, the left ventricle (LV), the myocardium (Myo) and the right ventricle (RV). For fair comparison, we conduct 2D segmentation experiments and use the same data split, the same data pre-processing and augmentations following (Luo et al., 2021). Specifically, we use 68 images (3/140 scans) and 136 images (7/140 scans) as our semi-supervised setting for training.

BraTS dataset includes 335 MR scans from four different modalities (FLARE, T1, T1ce and T2). Here we conduct 3D segmentation using the whole tumors from FLARE (250/335) similar to (Luo et al., 2022), utilizing the same data split, the same data augmentation and pre-processing strategy. For semi-supervised learning setting, we use 25/250 and 50/250 scans as different portions of the training sets.

AMOS dataset provides 500 CT and 100 MRI scans, each with voxel-level annotations of 15 abdominal organs such as Spleen, Left and Right kidney, etc. In this work, we only use 300 CT scans from the AMOS 2022 official grand challenge (Ji et al., 2022), including 200 scans for training and 100 scans for validation. In details, we divide the original validation set in a ratio of 3:7 to obtain the validation set and test set. Furthermore, 15/200 and 30/200 scans in the training set are used as two different semi-supervised settings.

During validation phase, we perform slice-to-slice inference for 2D segmentation and stack them into a 3D volume. For 3D segmentation, sliding window inference is applied to get the 3D predictions. To evaluate the segmentation results of both 2D and 3D, we use the metrics 1) Dice Coefficient (DSC) and 2) 95% Hausdorff Distance (HD95).

4.2. Implementation details

Network architectures and training details. Since we propose a plug-and-play scheme, we investigate the performance based on four different networks, including CNN-based and ViT-based networks for 2D and 3D segmentation, namely 2D-UNet (Ronneberger et al., 2015), SwinUNet (Cao et al., 2021), 3D-UNet and SwinUNETR (Hatamizadeh et al., 2022). Our experiments are implemented in Pytorch (Paszke et al., 2019) using an NVIDIA A100 GPU. All 2D networks are trained by an SGD optimizer for 30000 iteratoins, with an initial learning rate (lr) 0.01 and a batch size of 16. For 3D networks (trained on BraTS19), we use the same training strategy as 2D except the batch size is set to 4. Considering that the AMOS22 dataset with multiple categories may take longer time for the models to converge, we train the 3D networks (on AMOS22) for 60000 iterations with a batch size of 4 and an initial lr of 0.02. Specifically, during training, a batch of data consists of half labeled images and half unlabeled images. (More details in App. Sec.1)

Comparison with baselines and existing methods. We compare our proposed framework with four baselines and some recent state-of-the-art semi-supervised segmentation methods, including Entropy Minimization (Ent-Mini) (Vu et al., 2019), Cross Consistency Training (CCT) (Ouali et al., 2020), FixMatch (Sohn et al., 2020), Regularized Dropout (R-Drop) (Wu et al., 2021), Cross Pseudo Supervision (CPS) (Chen et al., 2021b), Uncertainty Rectified Pyramid Consistency (URPC) (Luo et al., 2022) and Cross Teaching between CNN and Transformer (CTCT) (Luo et al., 2021). For a fair comparison, all implementations of these methods are consistent with our framework under the same task. During validation, all the auxiliary training modules are discarded and only the trained backbones are used to produce final predictions. The above methods are open available (Luo, 2020).

4.3. Results of 2D segmentation on ACDC dataset

Improvements over baselines. Tab.1 and Tab.2 present the results of all methods on ACDC dataset using CNN-based and ViT-based backbones, respectively. Our proposed method surpasses the supervised baselines of CNN- and ViT-based backbones under 3 and 7 labeled scans by a large margin. Specifically, when using CNN-based backbone, the average improvements over the baselines are 28% in DSC and 43mm in HD95 under 3 labeled scans, 11.5% in DSC and 6.6mm in HD95 under 7 labeled scans. Especially for the class RV, our method improves the DSC from 41.26% to 81.75% and the HD95 from 79.78mm to 3.56mm when using only 3 labeled scans. As shown in Tab.2, our method still has promising improvements in DSC and HD95, yet some of the compared approaches suffer from declined performance when using ViT-based backbone.

Comparison with SOTA. When comparing the results with some current state-of-the-art methods, our approach outperforms the best under all partition protocols with both CNN- and ViT-based backbones, especially when the number of the labeled scans is extremely limited. In details, Tab.1 shows that our method suppasses the current SOTA method CTCT (Luo et al., 2021) by 8.21% in DSC and 14mm in HD95 under 3 labeled scans. In Tab.2, we surpasses CTCT using the same ViT backbone by 2.1% DSC and 2.44mm in HD95 under 7 labeled scans. We can also notice that our method has significant effects on the metric HD95, where we achieve 1.91mm and the fully supervised baseline is 1.56mm.

Fig. 3 presents the visual comparison between different CNN-based methods under 3 and 7 scans.

Table 1: Comparisons of our proposed methods with previous best studies on the ACDC using CNN-based backbone.

Labeled	CNN-Based	RV		Myo		LV		Mean	
		DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓
3 scans	UNet	41.26(33.74)	79.78(125.91)	56.73(26.44)	29.21(80.07)	65.04(29.25)	32.78(80.81)	54.34(29.81)	47.25(95.59)
	Ent-Mini	46.56(29.91)	39.86(32.72)	61.46(24.02)	29.37(67.14)	71.74(25.72)	33.11(68.69)	59.92(26.55)	34.11(56.18)
	CCT	46.88(31.96)	29.65(52.68)	63.18(22.92)	21.14(24.49)	68.88(26.47)	33.56(31.22)	59.65(27.12)	28.11(36.13)
	FixMatch	66.98(26.62)	28.22(67.68)	71.3(18.61)	9.22(21.52)	81.73(19.77)	9.03(20.07)	73.34(21.67)	15.49(36.42)
	R-Drop	43.55(34.04)	97.72(152.73)	63.46(25.02)	13.23(47.62)	72.93(25.95)	26.94(80.48)	59.98(28.34)	45.96(93.62)
	CPS	49.46(33.21)	43.93(90.65)	63.24(24.14)	18.9(66.64)	74.96(24.53)	28.01(67.59)	62.56(28.29)	30.28(74.96)
	URPC	51.05(27.98)	43.16(68.09)	60.43(24.13)	9.73(14.25)	71.32(26.53)	17.19(24.51)	60.93(26.21)	23.36(35.62)
	CTCT	71.68(26.24)	28.96(92.18)	70.58(23.82)	9.28(47.58)	80.23(23.56)	16.55(66.77)	74.16(24.54)	18.27(68.85)
	Ours	**81.75(10.82)**	**3.56(4.52)**	**79.37(13.09)**	**4.93(8.34)**	**86.01(17.46)**	**6.86(7.89)**	**82.37(13.79)**	**4.28(6.92)**
7 scans	UNet	68.92(28.24)	9.53(13.28)	78.26(9.24)	6.6(11.78)	85.18(11.36)	10.51(17.42)	77.46(16.28)	8.88(14.16)
	Ent-Mini	73.42(22.73)	4.85(7.98)	80.47(8.45)	5.36(10.19)	87.73(9.66)	9.14(16.76)	80.54(13.61)	6.45(11.64))
	CCT	80.08(15.95)	5.85(10.4)	82.85(7.63)	5.36(10.4)	89.17(9.13)	13.97(21.93)	84.01(10.9)	8.4(14.24)
	FixMatch	84.22(15.05)	2.01(2.50)	83.14(6.38)	2.5(5.73)	89.70(9.68)	6.08(12.67)	85.69(10.37)	3.53(6.97)
	R-Drop	72.7(26.22)	5.24(7.64)	81.38(6.79)	4.4(8.83)	89.28(8.42)	7.18(16.32)	81.12(13.81)	5.24(10.93)
	CPS	80.47(17.55	3.51(6.75)	82.65(6.38)	6.36(11.54)	87.89(10.03)	11.39(17.98)	83.67(11.32)	7.09(12.09)
	URPC	81.57(16.69)	4.3(7.69)	82.41(8.54)	5.95(13.12)	89.84(8.78)	8.31(17.76)	84.61(11.34)	6.19(12.86)
	CTCT	85.37(10.54)	3.45(8)	84.77(4.81)	7.05(18.04)	90.21(8)	9.2(16.2)	86.78(7.78)	5.43(10.58)
	Ours	**88.24(8.63)**	**1.67(1.46)**	**86.71(4.85)**	**1.6(1.94)**	**92(7.02)**	**3.54(6.88)**	**88.98(6.84)**	**2.27(3.43)**
140 scans (all)	UNet	90.07(7.53)	1.31(0.78)	88.87(3.3)	1.09(0.37)	94.32(3.8)	1.52(3.6)	91.09(4.88)	1.31(1.581)

Table 2: Comparisons of our proposed methods with previous best studies on the ACDC using ViT-based backbone.

Labeled	ViT-Based	RV		Myo		LV		Mean	
		DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓
3 scans	SwinUNet	58.53(27.06)	10.15(15.66)	58.69(21.84)	8.03(9.16)	71.63(24.86)	10.69(11.23)	62.95(24.58)	9.62(12.02)
	Ent-Mini	29.93(22.24)	16.9(11.82)	36.78(18.34)	20.01(16.31)	48.92(24.6)	27.53(19.72)	38.54(21.73)	21.48(15.95)
	R-Dropout	30.57(21.56)	24.82(21.27)	36.11(19.4)	18.31(16)	48.59(24.78)	23.67(15.33)	38.43(21.91)	22.27(17.53)
	CPS	30.51(21.58)	24.88(21.33)	36.13(19.41)	18.38(16.03)	48.63(24.76)	23.81(15.41)	38.43(21.92)	22.36(17.59)
	CTCT	72.99(22.93)	14.67(49.73)	70.75(20.14)	4.16(5.06)	81.15(20.52)	4.97(7.86)	74.96(21.2)	7.93(20.89)
	Ours	**76.85(14.21)**	**4.07(5.26)**	**72.31(16.36)**	**3.19(4.51)**	**85(14.42)**	**3.09(3.28)**	**78.06(15)**	**3.45(4.35)**
7 scans	SwinUNet	75.17(16.72)	10.74(21.86)	74.69(9.65)	4.75(6.9)	83.13(14.47)	9.6(13.36)	77.75(13.61)	8.36(14.04)
	Ent-Mini	44.15(22.17)	21.36(22.58)	50.46(17.72)	12.49(11.46)	62.64(21.32)	18.28(13.15)	52.42(20.4)	17.38(15.73)
	R-Dropout	48.42(20.77)	19.84(20.13)	49.58(17.46)	10.80(11.11)	63.81(21.9)	16.32(12.44)	53.94(20.04)	15.65(14.56)
	CPS	46.82(21.02)	23.64(22.35)	50.55(17.2)	11.8(11.59)	63.24(21.53)	17.40(12.38)	53.54(19.92)	17.61(15.44)
	CTCT	85.09(8.51)	2.86(7.98)	82.68(5.16)	2.69(4.97)	88.68(9.73)	7.51(13.74)	85.48(7.8)	4.35(8.9)
	Ours	**85.78(10.37)**	**1.86(1.63)**	**84.94(4.28)**	**1.8(4.57)**	**91.88(5.82)**	**2.08(5.63)**	**87.54(6.83)**	**1.91(3.94)**
140 scans (all)	SwinUNet	89.36(7.5)	1.82(3.9)	87.49(4.04)	1.16(0.75)	93.4(4.15)	1.69(3.13)	90.08(5.23)	1.56(2.59)

4.4. Results of 3D segmentation on BraTS and AMOS dataset

Improvements over baselines. For the BraTS Dataset, the quantitative comparisons are shown in Tab.4. It can be found that our method has significant improvements over the baselines under 25 and 50 annotated scans. Specifically, our CNN-based method outperforms the 3D-UNet with a gain of 2.23% in DSC and 3.24mm in HD95 under 25 labeled cases, and with a gain of 2.13% in DSC and 4.16mm in HD95 under 50 labeled cases. We can see that our method using only 20% labeled scans has even better results than the fully supervised baseline using 250 labeled scans in HD95, which is 7.44mm and 7.52mm, respectively.

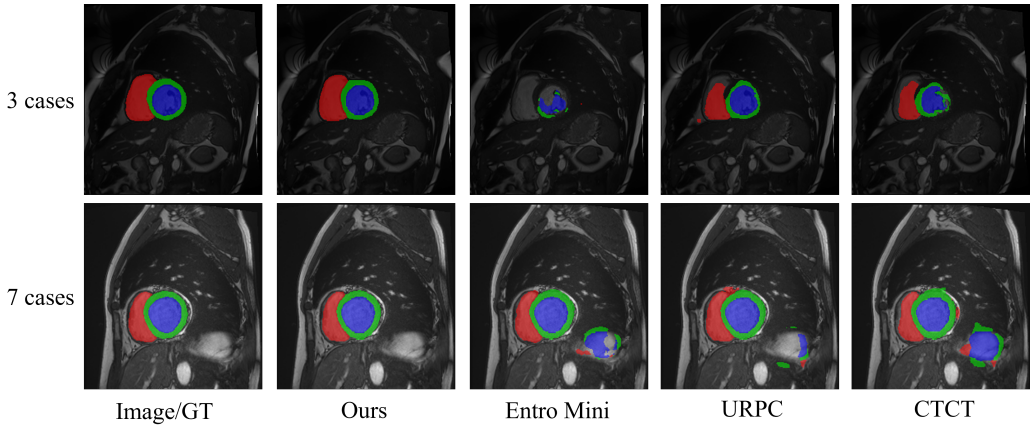

Figure 3: Visual comparison between different CNN-based methods under 3 and 7 scans. Red, Bule and Green masks denote the RV, LV and Myo respectively.

For the AMOS dataset, we perform experiments only on CNN-based backbone. The results are listed in Tab.5. As can be seen, our method exceeds the baselines by a large margin, especially under 15/200 protocol, achieving the improvement of over 40% in mean DSC and 90mm in mean HD95. When using 30 labeled scans, there is still a non-negligible improvement over the baseline method, with a gain of 14% in mean DSC and 12mm in mean HD95.

Comparison with SOTA. For the BraTS dataset, we compare our method with some recent semi-supervised segmentation methods. As observed, our proposed method outperforms all the compared methods with the highest DSC and lowest HD95 in all settings, including different partition protocols and backbones. In particular, we surpass the current best method URPC by a notable margin, achieving the improvements of +2.55mm and +2.31mm on the metric HD95 under 25/250 and 50/250 partition protocols respectively.

For the AMOS dataset, we conduct experiments on two existing methods considering the training cost and resource limitation. Tab.5 presents the performance of our methods and others under 15/200 and 30/200 labeled settings. Unsurprisingly, all of the methods reported significant improvements, comparing to partial supervised baseline, except for Entropy Mini under 30 labeled scans. From a comparison point of view, our proposed method achieves mean DSC of 61.44% and 74.81%, and mean HD95 of 17.85mm and 9.87mm on 15 labeled and 30 labeled cases, respectively, which is superior to other advanced methods, and its results are also very close to the full supervision results.

4.5. Ablation study:

In this section, we further conducted several experiments to investigate the effectiveness of each module in our framework.

Effects of CSAS and CSGS modules:

Effects of the consistency regularization part:

Table 3: Comparisons of our proposed methods with previous best studies on the BraTS using CNN-based and ViT-based backbone.

Network Architecture	Models	Tumor (25 scans)		Tumor (50 scans)		Tumor (250 scans)	
		DSC ↑	HD95 ↓	DSC ↑	HD95 ↓	DSC ↑	HD95 ↓
3D CNN-based	3D-UNet	82.67(12.8)	13.96(19.05)	83.76(12.27)	11.6(16.4)	86.77(11.14)	7.52(10.58)
	Ent-Mini	83.17(12.53)	12.01(17)	84.32(12.17)	10.59(14.81)		
	R-Dropout	83.21(11.46)	13.05(15.96)	84.41(11.7)	9.91(13.08)		
	CPS	83.62(11.6)	14.3(20.22)	84.26(12.43)	10.13(14.83)		
	CTCT	83.9(11.92)	12.32(18.24)	84.32(12.41)	10.01(14.72)		
	URPC	84.15(10.66)	13.27(19.4)	85.63(10.05)	9.75(13.1)		
	Ours	**84.9(10.23)**	**10.72(14.02)**	**85.89(10.16)**	**7.44(8.65)**		
3D ViT-based	SwinUNETR	83.9(12.3)	9.2(11.25)	84.87(11.57)	9.29(12.67)	87.33(10)	7.24(9.6)
	Ent-Mini	84.09(12.24)	10.67(13.84)	84.97(11.72)	8.29(10.28)		
	R-Dropout	84.05(11.76)	9.72(12.38)	84.83(10.91)	10.53(16.92)		
	CPS	83.79(12.52)	10.02(12.51)	84.99(11.16)	9.22(11.51)		
	CTCT	84.55(11.25)	10.21(13.12)	85.08(11.47)	9.65(13.17)		
	Ours	**85.38(10.71)**	**9.05(11.69)**	**85.63(10.93)**	**8.05(10.36)**		

Table 4: Comparisons of our proposed methods with previous best studies on the AMOS using CNN-based backbone.

		DSC ↑																
Labeled	CNN-Based	Large			Medium								Small					
		SPL	LIV	STO	RKI	LKI	BLA	AOR	PAN	IVC	DUO	PRO/UTE	GBL	ESO	RAG	LAG	Mean
15 scans	3D-UNet	57.92(21.14)	56.63(12.04)	10.93(9.6)	12.02(13.09)	6.38(8.74)	44.46(25.36)	53.02(17.92)	0.44(1.11)	37.19(17.38)	0.87(1.98)	32.81(29.64)	4.29(20.25)	2.18(5.33)	0.05(0.22)	0.25(1.93)	21.30(12.38)
	Ent-Mini	76.26(17.08)	80.94(9.41)	48.64(19.49)	73.42(19.24)	66.96(21.08)	42.08(21.34)	80.94(11.47)	30.61(16.9)	59.25(16.2)	15.84(14.44)	41.97(28.45)	23.32(22.74)	51.77(19.8)	29.78(20.19)	24.58(23.33)	49.76(18.74)
	CTCT	73.48(12.15)	70.13(10.07)	39.16(17.76)	73.45(19.11)	69.15(14.66)	27.79(20.76)	71.37(12.88)	22.49(14.31)	50.92(12.4)	13.72(12.91)	9.96(18.8)	4.29(20.25)	50.25(20.27)	13.04(14.89)	23.04(19.99)	40.82(16.08)
	Ours	88.11(12.38)	89.45(6.92)	60.29(22.28)	70.77(24.3)	79.87(19.35)	51.41(27.71)	87.15(9.81)	62.47(16.86)	69.98(11.99)	42.54(19.1)	34.87(26.5)	40.26(29.7)	59.46(19.15)	44.7(21.21)	40.21(25.21)	61.44(19.5)
30 scans	3D-UNet	78.9(12.21)	86.79(8.07)	61.16(23.53)	78.99(16.92)	82.08(15.7)	68.04(26.32)	77.42(16.42)	49.87(22.11)	65.59(17.06)	36.28(18.87)	63.79(23.15)	23.19(25.85)	54.81(21.55)	33.85(19.28)	48.36(20.03)	60.61(19.14)
	Ent-Mini	83.21(15.77)	90.07(8.44)	62.92(25.31)	81.66(18.01)	81.52(19.11)	52.02(28.86)	76.28(16.96)	41.45(22.45)	72.78(11.7)	42(20.39)	39.84(28.97)	41.64(30.19)	53.14(21.69)	47.83(22.61)	29.79(23.21)	59.74(20.91)
	CTCT	89.39(7.67)	93.32(5.51)	65.95(22.58)	90.59(12.68)	89.99(6.98)	55.21(23.32)	88.31(9.47)	65.58(14.33)	77.12(8.4)	52.03(15.96)	52.28(24.33)	63.38(25.03)	61.72(21.75)	64.74(13.96)	46.52(22.6)	70.41(15.64)
	Ours	81.63(15.2)	92.24(7.3)	75.01(20.85)	89.57(13.18)	91.22(10.12)	67.81(27.34)	90.77(7.33)	70.34(15.67)	83.78(6.84)	58.56(19.77)	66.78(21.84)	64.5(29.43)	70.4(17.08)	61.15(16.31)	58.44(20.12)	74.81(16.56)
200 scans	3D-UNet	93.92(7.12)	96.56(1.64)	86.94(17.51)	94.51(3.37)	92.36(12)	83.86(16.94)	93.5(5.36)	82.34(10.73)	88.57(4.47)	75.8(15.01)	77.76(18.08)	79.73(24.11)	76.57(14.88)	0	71.52(13.18)	79.6(10.96)
		HD95 ↓																
Labeled	CNN-Based	Large			Medium								Small					
		SPL	LIV	STO	RKI	LKI	BLA	AOR	PAN	IVC	DUO	PRO/UTE	GBL	ESO	RAG	LAG	Mean
15 scans	3D-UNet	29.86(34.87)	23.9(17.2)	65.03(52.7)	58.67(46.15)	67.65(83.24)	40.08(66.29)	13.99(10.52)	135.95(122.69)	18.07(10.88)	66.93(77.74)	36.42(78.67)	314.23(114.14)	75.33(98.28)	312.16(128.68)	357.7(72.93)	107.73(67.66)
	Ent-Mini	13.58(28.33)	13.89(24.21)	50.1(72.23)	12.52(45.61)	8.44(14.73)	17.57(43.32)	6.73(11.36)	21.29(43.16)	9.26(15.06)	26.44(43.58)	34.24(75.49)	55.83(86.59)	14.2(26.92)	18.6(47.28)	87.21(149.65)	25.99(48.5)
	CTCT	33.98(42.47)	11.11(11.52)	37.45(52.03)	33.98(42.47)	40.54(48.89)	25.17(43.8)	40.72(32.38)	36.52(31.28)	73.3(23.62)	39.49(45.24)	95.29(140.81)	357.14(75.57)	10.12(11.47)	108.05(161.82)	35.74(71.12)	63.98(56.06)
	Ours	9.13(23.78)	9.33(21.83)	22.49(47.53)	15.47(48.25)	6.43(16.48)	19.83(47.56)	3.05(5.99)	10.41(21.27)	6.44(8.38)	16.75(43.54)	32.97(74.74)	33.91(86.06)	5.25(5.63)	23.78(74.81)	52.56(123.2)	17.85(43.27)
30 scans	3D-UNet	69.54(40.23)	5.52(10.16)	22.62(45.18)	14.31(47.24)	16.22(31.63)	14.21(45.36)	7.21(10.19)	13.3(19.81)	6.67(6.61)	18.67(43.32)	17.3(61.34)	135.24(14.8)	14.8(44.49)	18.94(61.38)	15.04(61.48)	22.11(44.24)
	Ent-Mini	16.24(32.13)	7.87(18.29)	17.99(44.53)	4.29(12.13)	5.06(12.02)	16.5(43.88)	9.27(11.43)	15.73(17.97)	3.98(2.8)	17.96(43.63)	46.59(102.35)	31.86(84.70)	29.41(53.42)	15.31(61.48)	44.01(109.77)	18.8(43.37)
	CTCT	7.58(21.85)	4.61(17.33)	15.38(44.18)	6.98(44.24)	3.28(11.47)	16.6(44.04)	3.67(9.12)	7.25(7.04)	3.23(2.57)	13.96(43.52)	27.61(74.14)	23.45(76.08)	7.88(11.61)	2.81(2.1)	10.87(43.75)	10.34(30.2)
	Ours	7.77(15.91)	9.76(23.93)	13.5(44.51)	11.89(47.35)	8.07(21.22)	8.19(11.06)	2.11(5.04)	5.32(5.44)	1.91(1.29)	13.46(43.96)	20.87(64.23)	26.22(85.68)	6.67(13.37)	3.41(3.26)	8.89(43.91)	9.87(28.68)
200 scans	3D-UNet	1.11(4.43)	0.24(1.18)	8.79(44.49)	3.38(14.09)	1.09(3.56)	4.15(10.18)	1.26(3.87)	2.91(5.14)	1.29(0.62)	9.02(44)	15.09(61.91)	14.08(62.05)	3.1(4.56)	373.13(0)	2.36(2.96)	29.4(17.54)

5. Conclusion

Acknowledgments

Acknowledgments withheld.

References

Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27, 2014.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11): 2514–2525, 2018.

Table 5: CSAS Module on 2D Backbones (ACDC Dataset).

Labeled	2D Backbones	RV		Myo		LV		Mean	
		DSC	HD95	DSC	HD95	DSC	HD95	DSC	HD95
3 scans	UNet w/o CSAS	41.26(33.74)	79.78(125.91)	56.73(26.44)	29.21(80.07)	65.04(29.25)	32.78(80.81)	54.34(29.81)	47.25(95.59)
	UNet w/ CSAS	38.13(32.27)	104.38(149.63)	61.80(26.26)	22.71(67.03)	70.99(28.05)	20.14(51.33)	56.97(28.86)	49.08(89.33)
	SwinUNet w/o CSAS	58.53(27.06)	10.15(15.66)	58.69(21.84)	8.03(9.16)	71.63(24.86)	10.69(11.23)	62.95(24.58)	9.62(12.02)
	SwinUNet w/ CSAS	57.51(27.35)	7.38(13.27)	60.56(21.94)	8.5(9.86)	71.09(25.7)	11.17(11.49)	63.05(24.99)	9.01(11.54)
7 scans	UNet w/o CSAS	68.92(28.24)	9.53(13.28)	78.26(9.24)	6.6(11.78)	85.18(11.36)	10.51(17.42)	77.46(16.28)	8.88(14.16)
	UNet w/ CSAS	77.59(20.58)	5.9(9.71)	81.87(5.99)	8.08(15.89)	88.38(7.99)	11.16(18.94)	82.61(11.52)	8.38(14.85)
	SwinUNet w/o CSAS	75.17(16.72)	10.74(21.86)	74.69(9.65)	4.75(6.9)	83.13(14.47)	9.6(13.36)	77.75(13.61)	8.36(14.04)
	SwinUNet w/ CSAS	78.03(16.36)	2.46(2.11)	77.77(7.96)	4.17(6.55)	85.03(13.04)	7.39(10.28)	80.27(12.45)	4.67(6.31)
140 scans (all)	UNet w/o CSAS	90.07(7.53)	1.31(0.78)	88.87(3.3)	1.09(0.37)	94.32(3.8)	1.52(3.6)	91.09(4.88)	1.31(1.58)
	UNet w/ CSAS	90.53(7)	1.23(0.71)	89.04(3.34)	1.57(2.88)	94.27(4.07)	1.04(0.19)	91.28(4.81)	1.28(1.26)
	SwinUNet w/o CSAS	89.36(7.5)	1.82(3.9)	87.49(4.04)	1.16(0.75)	93.4(4.15)	1.69(3.13)	90.08(5.23)	1.56(2.59)
	SwinUNet w/ CSAS	90.13(7.32)	1.24(0.63)	88.27(3.22)	1.04(0.12)	93.89(4)	1.08(0.27)	90.76(4.84)	1.12(0.33)

Table 6: CSAS Module on 3D Backbones (AMOS Dataset).

Labeled	3D Backbone	Metric DSC ↑															
		Mean	SPL	RKI	LKI	GBL	ESO	LIV	STO	AOR	IVC	PAN	RAG	LAG	DUO	BLA	PRO/UTE
15 scans	3D-UNet w/o CSAS	21.30(12.38)	57.92(21.14)	12.02(13.09)	6.38(8.74)	4.29(20.25)	2.18(5.33)	56.63(12.04)	10.93(9.6)	53.02(17.92)	37.19(17.38)	0.44(1.11)	0.05(0.22)	0.25(1.93)	0.87(1.98)	44.46(25.36)	32.81(29.64)
	3D-UNet w/ CSAS	46.71(20.58)	67.76(23.72)	47.43(27.26)	70.4(21.93)	36.01(27.48)	54.08(21.41)	76.79(12.48)	37.25(23.38)	74.23(17.23)	72.6(10.71)	44.21(19.35)	35.52(22.66)	22.54(23.59)	30.08(18.46)	15.83(19.14)	15.98(19.92)
30 scans	3D-UNet w/o CSAS	60.61(19.14)	78.9(12.21)	78.99(16.92)	82.08(15.7)	23.19(25.85)	54.81(21.55)	86.79(8.07)	61.16(23.53)	77.42(16.42)	65.59(17.06)	49.87(22.11)	33.85(19.28)	48.36(20.03)	36.28(18.87)	68.04(26.32)	63.79(23.15)
	3D-UNet w/ CSAS	72.31(17.41)	83.56(15.42)	92.6(7)	89.19(13.2)	61.1(29.69)	69.24(17.28)	90.42(9.28)	72.24(22.23)	88.77(9.3)	82.1(7.24)	66.08(18.38)	62.05(16.77)	51.23(23.66)	44.78(19.08)	70.35(25.81)	60.89(26.81)
200 scans (all)	3D-UNet w/ CSAS	79.6(10.96)	93.92(7.12)	94.51(3.37)	92.36(12)	79.73(24.11)	76.57(14.88)	96.56(1.64)	86.94(17.51)	93.5(5.36)	88.57(4.47)	82.34(10.73)	0	71.52(13.18)	75.8(15.01)	83.86(16.94)	77.76(18.08)
	3D-UNet w/o CSAS	84.92(11.09)	94.18(6.73)	95.25(2.58)	94.85(6.25)	81.62(20.8)	76.15(14.77)	96.77(1.44)	86.84(17.51)	93.18(4.98)	88.61(4.65)	83.08(11.44)	68.63(11.99)	70.98(11.66)	76.77(15.66)	84.76(17.91)	82.06(17.94)

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021a.

Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021b.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2022.

Yuanfeng Ji, Haotian Bai, Jie Yang, Chongjian Ge, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *arXiv preprint arXiv:2206.08023*, 2022. URL https://amos22.grand-challenge.org/.

Rushi Jiao, Yichi Zhang, Le Ding, Rong Cai, and Jicong Zhang. Learning with limited annotations: a survey on deep semi-supervised learning for medical image segmentation. *arXiv preprint arXiv:2207.14191*, 2022.

Xiangde Luo. 2020. URL https://github.com/HiLab-git/SSL4MIS.

Table 7: CSAS Module on 3D Backbones (AMOS Dataset).

Labeled	3D Backbone	Metric HD95 ↓															
		Mean	SPL	RKI	LKI	GBL	ESO	LIV	STO	AOR	IVC	PAN	RAG	LAG	DUO	BLA	PRO/UTE
15 scans	3D-UNet w/o CSAS	107.73(67.66)	29.86(34.87)	58.67(46.15)	67.65(83.24)	314.23(114.14)	75.33(98.28)	23.9(17.2)	65.03(52.7)	13.99(10.52)	18.07(10.88)	135.95(122.69)	312.16(128.68)	357.7(72.93)	66.93(77.74)	40.08(66.29)	36.42(78.67)
	3D-UNet w/ CSAS	34.77(57.08)	30.57(39.29)	23.41(48.61)	16.69(26.61)	48.01(109.34)	14.72(19.3)	27.05(33.19)	31.97(48.5)	8.44(11.34)	14.77(24.94)	13.27(19.24)	22.29(74.36)	112.28(165.02)	21.66(44.6)	73.39(101.12)	63.02(90.67)
30 scans	3D-UNet w/o CSAS	22.11(44.24)	69.54(40.23)	14.31(47.24)	16.22(31.63)	135.24(14.8)	14.8(44.49)	5.52(10.16)	22.62(45.18)	7.21(10.19)	6.67(6.61)	13.3(19.81)	18.94(61.38)	15.04(61.48)	18.67(43.32)	14.21(45.36)	17.3(61.34)
	3D-UNet w/ CSAS	10.96(24.6)	36.16(46.47)	1.12(2.24)	1.6(2.68)	28.21(85.81)	6.17(11.38)	9.17(22.99)	15.21(44.6)	3.02(6.82)	3.14(6.59)	6.44(6.9)	3.18(3.05)	4.76(3.75)	15.96(43.63)	10.53(20.6)	19.76(61.44)
200 scans (all)	3D-UNet w/ CSAS	29.4(17.54)	1.11(4.43)	3.38(14.09)	1.09(3.56)	14.08(62.05)	3.1(4.56)	0.24(1.18)	8.79(44.49)	1.26(3.87)	1.29(0.62)	2.91(5.14)	373.13(0)	2.36(2.96)	9.02(44)	4.15(10.18)	15.09(61.91)
	3D-UNet w/ CSAS	4.35(17.27)	1.32(5.19)	1.48(9.9)	0.51(1.88)	13.18(61.98)	3.79(6.01)	0.17(0.77)	9.33(44.45)	1.37(3.9)	1.38(0.86)	2.44(4.03)	2.21(1.79)	2.37(1.91)	8.76(44)	4.04(10.5)	12.91(61.87)

Xiangde Luo, Minhao Hu, Tao Song, Guotai Wang, and Shaoting Zhang. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. *arXiv preprint arXiv:2112.04894*, 2021.

Xiangde Luo, Guotai Wang, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Dimitris N Metaxas, and Shaoting Zhang. Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *Medical Image Analysis*, 80: 102517, 2022.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Jizong Peng, Guillermo Estrada, Marco Pedersoli, and Christian Desrosiers. Deep co-training for semi-supervised image segmentation. *Pattern Recognition*, 107:107269, 2020.

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

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021.

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019.

Ping Wang, Jizong Peng, Marco Pedersoli, Yuanfeng Zhou, Caiming Zhang, and Christian Desrosiers. Self-paced and self-consistent co-training for semi-supervised image segmentation. *Medical Image Analysis*, 73:102146, 2021.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905, 2021.

Enze Xie, Wenjia Wang, Wenhai Wang, Peize Sun, Hang Xu, Ding Liang, and Ping Luo. Segmenting transparent object in the wild with transformer. *arXiv preprint arXiv:2101.08461*, 2021.

Yuyin Zhou, Yan Wang, Peng Tang, Song Bai, Wei Shen, Elliot Fishman, and Alan Yuille. Semi-supervised 3d abdominal multi-organ segmentation via deep multi-planar co-training. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 121–140. IEEE, 2019.

Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*, 2020.

Appendix A. Appendix 1