COMPUTATIONALLY-EFFICIENT VISION TRANSFORMER FOR MEDICAL IMAGE SEMANTIC SEGMENTATION VIA DUAL PSEUDO-LABEL SUPERVISION

Ziyang Wang, Nanqing Dong, Irina Voiculescu

Department of Computer Science, University of Oxford, UK

ABSTRACT

Ubiquitous accumulation of large volumes of data, and increased availability of annotated medical data in particular, has made it possible to show the many and varied benefits of deep learning to the semantic segmentation of medical images. Nevertheless, data access and annotation come at a high cost in clinician time. The power of Vision Transformer (ViT) is well-documented for generic computer vision tasks involving millions of images of every day objects, of which only relatively few have been annotated. Its translation to relatively more modest (i.e. thousands of images of) medical data is not immediately straightforward. This paper presents practical avenues for training a **C**omputationally-**E**fficient **S**emi-**S**upervised **V**ision **T**ransformer (CESS-ViT) for medical image segmentation task.

We propose a pure self-attention-based image segmentation network which requires only limited computational resources. Additionally, we develop a dual pseudo-label supervision scheme for use with semi-supervision in a ViT.

Our method has been evaluated on a publicly available cardiac MRI dataset with direct comparison against other semi-supervised methods. Our results illustrate the proposed ViT-based semi-supervised method outperforms the existing methods in the semantic segmentation of cardiac ventricles.

Index Terms— Semantic Segmentation, Semi-Supervised Learning, Vision Transformer, Cardiac MRI

1. INTRODUCTION

Due to the high cost of clinician annotation and ready availability of unannotated medical scan data, semi-supervised learning (SSL) has been gaining momentum in the semantic segmentation of medical images. The principal strength of SSL is that it trains on a mixture of large amounts of unlabeled data and a small amount of labeled data for the training model. SSL approaches such as [1] on skin lesions have started to emerge. This relied on a weighted sum of supervision loss and regularization loss for labeled and unlabeled data. Yu [2] proposed a semi-supervised uncertainty aware framework for 3D MRI images. In the context of more generic computer vision tasks, the Mean Teacher [3] or Temporal Ensembling [4] have been successful uncertainty-aware schemes which gradually learn from reliable targets. Luo further explored uncertainty rectified pyramid consistency regularization, and pixel-level classification task was also explored by [5][6].

More recent challenges (albeit in natural language processing) are tackled by Transformers using self-attention to model long-range dependencies [7]. Vision Transformer (ViT), a self-attention-based architecture for image classification tasks, was firstly proposed in [8]. Variants such as the Swin-Transformer [9] use a shift window and patch merging to detect large variations in the scale of visual entities, with efficient computational cost for high resolution images; DieT [10] introduces a new distillation procedure to train with lower computation resources that ensure a 'student' learning from a 'teacher' using attention; TransUNet [11] combines U-Net and ViT; TNT [12] considers image patches as 'visual sentences' and 'visual words'. Training ViT in a SSL manner remains a challenging, under-studied topic, with good potential for medical imaging [13].

2. MOTIVATION

Our work has been driven by the unexplored usage of ViT for medical imaging applications, where the volumes of unlabeled training data are several orders of magnitude smaller than in conventional computer vision tasks. Our ideal scenario is to achieve segmentation quality similar to that already available through the use of CNN, but for a fraction of the image annotations and in a computationally efficient manner.

To this end, we introduce a semi-supervised ViT via dual pseudo-label supervision. Our methodology contributes to the field on two accounts:

1) We explore a pure self-attention-based ViT for medical image semantic segmentation. This is tested under limited computational resources, comparing its approximate training time costs against conventional CNN [14].

2) We develop a *dual* pseudo-label supervision framework for ViT as detailed in the remainder of this paper.

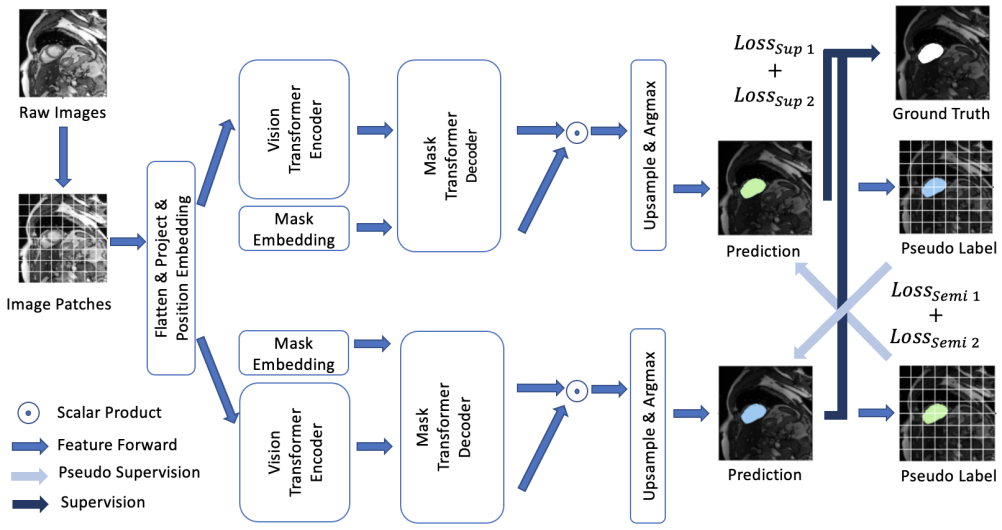

Fig. 1: Framework for Computationally-Efficient Semi-Supervised Vision Transformer for Medical Image Segmentation

3. METHOD

3.1. Outline

The architecture of our framework, sketched in Figure 1 and detailed in Section 3.2, is a *dual* pseudo-label supervision framework for ViT. Two ViT encoder-decoders with identical architectures are employed for dual-view learning; pseudo label are generated in order to allow for expanding the data set with unlabeled data. One ViT's inference is then used as the other ViT's training labels. The process is then repeated with the roles reversed. Inspired by consistency regularization [15], one ViT can supervise the other's feature learning. Two different initializations of the ViTs, using the same input raw data, will cause non-identical but similar output inferences. This scheme significantly outperforms other existing CNN-based SSL, especially using limited amounts of annotated data.

3.2. Vision Transformer for Semantic Segmentation

Semantic features are essential in dense recognition tasks such as semantic segmentation, image detection. CNN-based networks are normally developed with multi-layers of convolutional layers with limited receptive field, up/down sampling layers, and a large number of channels to transfer high-level feature information. The final layer is normally set as Softmax with a dense map. In this way, a pixel-level classification tasks can be achieved [16]. The image feature information, however, is going to be blurred after multi-layers encoders [17]. In U-Net [14], copy and crop is utilized between encoder and decoder to make sufficient semantic feature information been transferred through CNN which results in a dominant position in the segmentation tasks.

A fundamental limitation of CNNs is that modeling global dependencies is still the barrier to improving semantic segmentation performance. By contrast, our pure self-attention ViT for sequence to sequence for semantic segmentation models long-range dependencies. Modeling long-range dependencies often requires high computational cost and training time. Our proposed ViT-based network is not only copes well with the segmentation task, but also achieves training costs similar to those of a conventional U-Net.

3.3. Semi-Supervised Learning Setup

In a generic SSL image segmentation task, $\mathbf{L}$, $\mathbf{U}$ and $\mathbf{T}$ normally denote a small labeled dataset, a large size of unlabeled dataset, and a dataset for testing. In our baseline experiment they are set, respectively, to 10%, 70% and 20% of the dataset, with no overlap. Further experimentsin Section 4 make different assumptions about these ratios.

We denote a batch of labeled data as $(\boldsymbol{X}_{\mathrm{l}}, \boldsymbol{Y}_{\mathrm{gt}}) \in \mathbf{L}$, $(\boldsymbol{X}_{\mathrm{t}}, \boldsymbol{Y}_{\mathrm{gt}}) \in \mathbf{T}$ for labeled training and testing data with its corresponding ground truth, and a batch of only raw data as $(\boldsymbol{X}_{\mathrm{u}}) \in \mathbf{U}$ in the unlabeled dataset, where $\boldsymbol{X} \in \mathbb{R}^{h \times w}$ representing a 2D image. $\boldsymbol{Y}_{\mathrm{p}}$ is the dense map predicted by a segmentation transformer $f_{\mathrm{t}} : \boldsymbol{X} \mapsto \boldsymbol{Y}_{\mathrm{p}}$ as pseudo label for training. Final evaluation results are calculated based on comparisons between Y_{p} and Y_{gt}.

Inspired by Transformer [7], ViT [8], DETR [18], and Segmentor [19], our proposed ViT-based segmentation network consists of a self-attention-based encoder-decoder for segmentation. A sequence of patches $\boldsymbol{X}' = [x'_1 \cdots x'_N]^\top \in \mathbb{R}^{N \times P^2}$ is sourced from an input raw 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 flattened into a 1D vector and projected with patch embedding

$\boldsymbol{X}_0 = [E_1 \cdots E_N]^\top, E_{1\cdots N} \in \mathbb{R}^{D\times P^2}$. The positional embeddings to provide the positional information of each patch $pos = [pos_1 \cdots pos_N]^\top \in \mathbb{R}^{N\times D}$ 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 multi-layer perceptron (MLP) block of two layers [7]. Residual connections and layer normalization (LN) are both applied in each block [20][21]. The details of MSA and MLP two sublayers are calculated as $\boldsymbol{A}_{i-1} = \mathrm{MSA}(\mathrm{LN}(\boldsymbol{Z}_{i-1})) + \boldsymbol{Z}_{i-1}$, $\boldsymbol{Z}_i = \mathrm{MLP}(\mathrm{LN}(\boldsymbol{A}_{i-1})) + \boldsymbol{A}_{i-1}$, where $i \in 1, \ldots, L$, and L is the number of identical 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 $\boldsymbol{V}$ [7], which is calculated $\boldsymbol{Q} = \mathrm{Linear_Q}(\boldsymbol{Z}'), \boldsymbol{K} = \mathrm{Linear_K}(\boldsymbol{Z}'), \boldsymbol{V} = \mathrm{Linear_V}(\boldsymbol{Z}')$. In this way, the transformer encoder maps input sequence $\boldsymbol{Z}_0 = [z_{0,1} \cdots z_{0,N}]$ with position to $\boldsymbol{Z}_L = [z_{L,1}, ..., z_{L,N}]$.

The above settings are adapted from [8], and a proper setting of efficient computational cost is designed. In the transformer decoder, the sequence of Z_L is then decoded to a dense map $\boldsymbol{S} \in \mathbb{R}^{h\times w\times k}$ as segmentation results, where k is the number of classes. The decoder acts as mapping patch from the encoder and unsample to pixel-level probability of dense map. 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 up-sampled to the size of input raw image as prediction results. Both class embedding and patch sequence are jointly processed, and semantic segmentation mask is finally generated.

After a series of experiments of implementing segmentation ViT with semi-supervised training, and following the goal of achieving competitive segmentation performance under similar training cost as CNN (i.e. U-Net), we empirically determined the patch size to 16×16, the number of multi-heads in self-attention sublayer of 6, the number of identical layers L in the encoder as 12, the number of identical layers M of the decoder as 2, and the dimension of the output from each self-attention layer as 384.

3.4. Semi-Supervised Dual Pseudo-Label Supervision

Due to the high human cost of data annotation, our goal is to achieve good quality results using small amounts of labeled data but potentially making use of relatively larger amounts of unlabeled data, increasingly available in all clinics. Encouraging the consistency of prediction by adding perturbations, also known as consistency regularization, is an essential approach in SSL [22]. For example, [1] introduced perturbation through adding transformations into the training data; [2] and [3] added dropout into network as perturbation to improve semi-supervised learning performance.

Inspired by [23][15][13] we introduce a dual pseudo-label supervision semi-supervised framework for ViT. Two segmentation ViTs f_t with the same architecture are initialized with different weight parameters. The output from the same input image $\boldsymbol{X}$ can be illustrated as $\boldsymbol{Y}_\mathrm{p1} = argmax(f_\mathrm{t}(\boldsymbol{X}; \theta_1))$ and $\boldsymbol{Y}_\mathrm{p2} = argmax(f_\mathrm{t}(\boldsymbol{X}; \theta_2))$ where θ is the weight set of each network. The inference from one ViT is input as pseudo labels into the training of the other ViT.

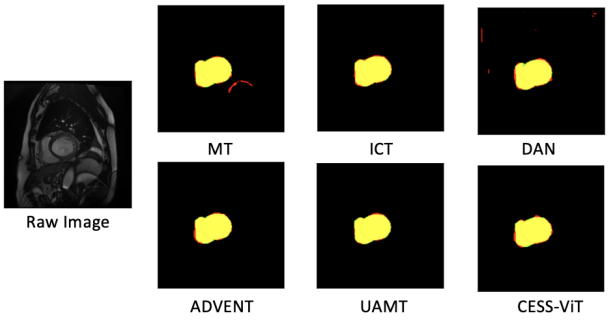

Fig. 2: An Example Raw Image and Inference Results on Testing Set

The training objective is to minimize the sum of semi-supervision loss $Loss_\mathrm{semi}$ and supervision loss $Loss_\mathrm{sup}$ of the two ViTs f_t. The semi-supervision loss for each ViT $Loss_\mathrm{semi1}$ or $Loss_\mathrm{semi2}$ are calculated using cross-entropy CE as shown in Equation 1:

$$Loss_\mathrm{semi1or2} = \mathrm{CE}\big(argmax(f_\mathrm{t}(\boldsymbol{X}; \theta_\mathrm{1or2}), f_\mathrm{t}(\boldsymbol{X}; \theta_\mathrm{2or1}))\big) \tag{1}$$

The supervision loss for each ViT is calculated using both CE and the Dice Coefficient $Dice$ as shown in Equation 2:

$$\begin{aligned} Loss_\mathrm{sup1or2} = \frac{1}{2} \times \big(&\mathrm{CE}(Y_\mathrm{gt}, f_\mathrm{t}(\boldsymbol{X}; \theta_\mathrm{2or1})) + \\ &Dice(Y_\mathrm{gt}, f_\mathrm{t}(\boldsymbol{X}; \theta_\mathrm{2or1}))\big) \end{aligned} \tag{2}$$

The overall loss of CESS-ViT being optimized during training is detailed in Equation 3:

$$Loss = Loss_\mathrm{sup1} + Loss_\mathrm{sup2} + \lambda(Loss_\mathrm{semi1} + Loss_\mathrm{semi2}) \tag{3}$$

where λ is the weight for consistency loss, and it is updated every 150 iterations. It is a trade-off weight that keeps increasing during training process following by [4].

4. EXPERIMENTS

4.1. Dataset and Experimental Setup

Our experiments validate the method on the MRI ventricle segmentation dataset from the automated cardiac diagnosis MICCAI Challenge 2017 [14]. There is data from 100 patients (nearly 6 000 images) covering different distributions of feature information, across five evenly distributed subgroups:

Label/Total	10%				
Model	Dice	Acc	Pre	Sen	Spe
MT[3]	0.8593	0.9887	0.8576	0.8610	0.9941
ICT[24]	0.8972	0.9919	0.9125	0.8823	0.9965
DAN[25]	0.5395	0.9480	0.4172	0.7631	0.9557
ADVENT[26]	0.8612	0.9896	**0.9258**	0.8051	**0.9973**
UAMT[2]	0.8347	0.9873	0.8683	0.8037	0.9949
DCN[27]	0.8787	0.9908	0.9248	0.8370	0.9972
CESS-ViT (ours)	**0.9034**	**0.9923**	0.9066	**0.9002**	0.9961

Table 1: The Direct Comparison with Similarity Measures on Cardiac MRI Testing Set (the higher, the better)

Label/Total	10%		
CESS-ViT	RVD	HD	ASSD
MT[3]	0.3715	28.5797	6.4947
DAN [25]	0.2259	145.4982	49.5673
ICT[24]	**0.2063**	22.4767	4.4015
ADVENT[26]	0.2669	20.3860	4.7762
UAMT[2]	0.3925	27.2209	6.4702
DCN[27]	0.2630	21.0363	4.3865
CESS-ViT (ours)	0.2315	**10.4456**	**2.7204**

Table 2: The Direct Comparison with Difference Measures on Cardiac MRI Testing Set (the lower, the better)

Label/Total	2%	5%	10%	15%	20%
MT[3]	0.2544	0.6241	0.8593	0.9103	0.9137
UAMT[2]	0.2335	0.3975	0.8347	0.8999	0.8907
DCN[27]	0.2194	0.5497	0.8787	0.8883	0.9050
CESS-ViT (ours)	**0.7396**	**0.8350**	**0.9034**	**0.9129**	**0.9164**

Table 3: The Dice Results under Different Assumption of Percentage of Label/Total Data (the higher, the better)

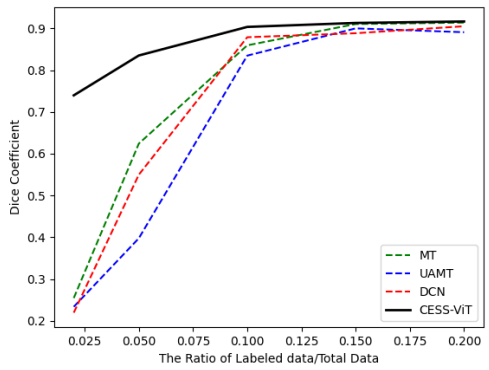

Fig. 3: The Dice Results under Different Assumption of the Percentage of Label/Total Data

normal, myocardial infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy, and abnormal right ventricle.

Our code has been developed under Ubuntu 20.04 in Python 3.8.8 using Pytorch 1.10 and CUDA 11.3 using four Nvidia GeForce RTX 3090 GPU with 24GB memory, and Intel(R) Intel Core i9-10900K. The ViT segmentation backbone is based on [19]. The runtimes averaged around 6 hours, including the data transfer, training, inference and evaluation. Dataset is processed for 2D image segmentation purpose, and all images are resized to 256×256. All algorithms including CESS-ViT and other baseline methods are trained with same hyperparameter setting(no modification of CSEE-ViT) including: 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.

4.2. Results and Discussion

CESS-ViT is compared with other CNN-based SSL: MT [3], ICT [24], DAN [25], ADVENT [26], UAMT [2], DCN [27] with a U-Net backbone [28].

Figure 2 illustrates a sample raw image with related predicted images against the published ground truth. Yellow, red, green and black show true positive (TP), false positive (FP), false negative (FN) and true negative (TN) pixels. This illustrates how CESS-ViT can give rise to fewer FP pixels and lower Hausdorff Distance (HD) compared to other methods.

Quantitative comparisons are conducted with a variety of evaluation metrics including similarity measures: Dice, Accuracy, Precision, Recall/Sensitivity, Specificity. We also investigate difference measures: Relative Volume Difference (RVD), HD, Average Symmetric Surface Distance (ASSD). The results reported in Tables 1 and 2 are conducted under the assumption that 10% of the training set has been labeled. Table 3 gives a systematic review of how the Dice coefficient varies when only 2%, 5%, 10%, 15%, and 20% of the training set is labeled. Figure 3 illustrates how the different ratios of labeled data to total data influence the segmentation results: whilst all methods cope reasonably with about 15% of labeled images, our method yelds usable results at very low percentages of data having been labeled. This gives hope for leveraging the vast volumes of data from clinics.

5. CONCLUSIONS

It is pleasantly remarkable to see serviceable results being obtained with a proportion of labelled data as small as 2% of the total. This opens an important avenue for exploiting large unlabeled datasets which clinicians have only limited time to annotate. Conversely, given small existing collections of annotations of a particular kind, these can now be put to good use through pairing them up with large volumes of unlabeled data and feeding them through framewoks such as CESS-ViT.

6. REFERENCES

[1] Xiaomeng Li and etc, "Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model," *BMVC*, 2018.

[2] Lequan Yu and etc, "Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation," in *MICCAI*. Springer, 2019.

[3] Antti Tarvainen and etc, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *NIPS*, 2017.

[4] Samuli Laine and Timo Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

[5] Xiangde Luo and etc, "Semi-supervised medical image segmentation through dual-task consistency," in *AAAI Conference on Artificial Intelligence*, 2021, pp. 8801–8809.

[6] Xiangde Luo and etc, "Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency," in *MICCAI*, 2021.

[7] Ashish Vaswani and etc, "Attention is all you need," in *Advances in neural information processing systems*, 2017.

[8] Alexey Dosovitskiy and etc, "An image is worth 16x16 words: Transformers for image recognition at scale," *International Conference on Learning Representations*, 2021.

[9] Ze Liu and etc, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022.

[10] Hugo Touvron and etc, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10347–10357.

[11] Jieneng Chen and etc, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[12] Kai Han and etc, "Transformer in transformer," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[13] Xiangde Luo and etc, "Semi-supervised medical image segmentation via cross teaching between cnn and transformer," *arXiv preprint arXiv:2112.04894*, 2021.

[14] Olivier Bernard and etc, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?," *IEEE transactions on medical imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[15] Xiaokang Chen and etc, "Semi-supervised semantic segmentation with cross pseudo supervision," in *CVPR*, 2021.

[16] Ziyang Wang and etc, "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)*. IEEE, 2021.

[17] Zhengdong Zhang and etc, "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)*. IEEE, 2020.

[18] Nicolas Carion and etc, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[19] Robin Strudel and etc, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7262–7272.

[20] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[21] Kaiming He and etc, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[22] 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*, 2020.

[23] Zhanghan Ke and etc, "Dual student: Breaking the limits of the teacher in semi-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6728–6736.

[24] Vikas Verma and etc, "Interpolation consistency training for semi-supervised learning," in *International Joint Conference on Artificial Intelligence*, 2019.

[25] Yizhe Zhang and etc, "Deep adversarial networks for biomedical image segmentation utilizing unannotated images," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017.

[26] Tuan-Hung Vu and etc, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *CVPR*, 2019.

[27] Siyuan Qiao and etc, "Deep co-training for semi-supervised image recognition," in *ECCV*, 2018.

[28] Xiangde Luo, "SSL4MIS," `https://github.com/HiLab-git/SSL4MIS`, 2020.