Adversarial Vision Transformer for Medical Image Semantic Segmentation with Limited Annotations

Ziyang Wang[1]
ziyang.wang@cs.ox.ac.uk
Chengkuan Zhao[2]
cz015@bucknell.edu
Zixuan Ni[3]
zini8743@colorado.edu

[1] Department of Computer Science
University of Oxford
[2] Department of Computer Science
Bucknell University
[3] Department of Computer Science
University of Colorado Boulder

Abstract

Medical image analysis has benefited from deep learning techniques not only because of network architecture engineering, but also a large number of high-quality annotations which is time- and labour-consuming. Motivated by the recent success of Vision Transformer(ViT), we propose to explore the power of ViT for medical image semantic segmentation with an advanced Semi-Supervised Learning(SSL) fashion via *MixUp-based interpolation consistency* training and *adversarial* training. Aiming to train Segmentation ViT model(sViT) with labelled and unlabelled data simultaneously, an adversarial SSL framework that consists of a sViT and an evaluation model(EM) is proposed in this paper. During the adversarial training process, the EM is trained to classify the quality of inference of sViT is from labelled/unlabelled sample, and the sViT is initialized and trained against EM (i.e. all inference by sViT is high-quality enough to be classified as if from labelled data). To further boost the performance of sViT, MixUp-based interpolation consistency training is introduced and utilized for sViT. The whole adversarial training is designed separately for sViT and EM in an iterative manner, and the MixUp is solely for sViT. Experimental results(including replacing sViT to CNN) demonstrate the proposed method competitive performance against other SSL methods on a public benchmark data set with a variety of metrics. The code is publicly available on GitHub.[1]

1 Introduction

Deep-learning-based methods, especially Convolutional Neural Networks (CNN), have dominated in medical image semantic segmentation since 2015 when FCN was proposed [26]. It is the first fully CNN-based network trained end-to-end and pixels-to-pixels. To tackle the lack of semantic feature information being transferred through multi-CNN layers, a skip connection hereby is proposed with CNN results in one of the most promising semantic segmentation backbone networks i.e. UNet [11]. Except for the backbone network study such as UNet, Deeplab[7], and PSPNet[53], many of the studies also intent to explore a variety of

[1] https://github.com/ziyangwang007/CV-SSL-MIS

advanced network blocks such as multi-scale learning[23, 43], attention mechanism[12, 14], feature normalization[18], residual learning[9], densely connected[15], and atrous/dilated CNN[49] for the backbone which results in a family of UNet-based segmentation methods for CT, Ultrasound, and MRI [10, 17, 19, 46].

The most recent study, however, demonstrates that Transformer, a pure self-attention-based method, outperforms RNN-based methods with different blocks in the natural language process [38]. The ViT is then proposed to explore the feature learning ability of Transformer in the computer vision domain [8]. A family of ViT-based methods for dense prediction tasks is then presented such as Swin-ViT to use shift window for hierarchical ViT results in state-of-the-art results in 2021 [24]. Some other ViT segmentation backbone studies are motivated by Encoder-Decoder style segmentation models such as TransUNet[6], which presents the ViT improve conventional UNet in the encoder; and Swin-UNet[4], which directly presents the Swin-ViT blocks applied into U-shape backbone network.

Although the recent study of network architecture engineering with ViT demonstrate promising performance in the computer vision community, the training strategy is still one of the essential barriers to deploying advanced techniques in clinical medical image analysis. In the medical imaging community, medical data normally comes up with a large amount of raw data with a small proportion of annotations due to the high cost of clinician annotation with the requirement of high expert level skill, SSL has been gaining momentum in medical image segmentation. SSL allows the model to be beneficial with labelled and unlabelled data together but remains challenging with ViT [27, 41, 42].

Following the above concern, we propose to explore a simple and bespoke sViT with the SSL fashion for medical image segmentation. A MixUp-based **C**onsistency-**A**ware **A**dversarial **Vi**sion **T**ransformer, called CAA-ViT, is presented. The contribution of CAA-ViT is fourfold as follows:

- We propose to use CNN or Swin-ViT network blocks directly with the U-shape Encoder-Decoder style segmentation network architecture, respectively.
- Two adversarial SSL training stages are proposed, which consist of training a ViT-based segmentation model(sViT), and a CNN-based evaluation model(EM), respectively.
- To boost the performance of sViT, a MixUp-based interpolation consistency training is proposed under consistency-aware concern.
- The CAA-ViT is evaluated on a public benchmark data set with a variety of evaluation measures, and outperforms other semi-supervised methods [32, 36, 39, 40, 41, 42, 50, 52] under the same setting and feature information distribution.

2 Related Work

2.1 Segmentation Vision Transformer

Image semantic segmentation can be considered a dense prediction task to classify each pixel of an image whether belonging to a region of interest or background. It is one of the most challenging computer vision tasks and has been dominated by CNN since 2015 before the emergence of ViT [11, 19, 26]. CNN-based methods, however, are limited to its finite receptive field [7, 38, 49]. Motivated by the recent success of self-attention in natural language

processing [38], many image processing tasks are also explored as sequence-to-sequence problems and the Vision Transformer(ViT) firstly demonstrates the promising performance of self-attention in image recognition [8]. The input image is split as a series of image patches with location information, and hereby the global dependencies can be modeled with pure multi-head self-attention. A variety of ViT-based convolution-free backbone network for dense prediction are widely proposed such as shift-window ViT(Swin-ViT)[24], and sample efficient ViT(DeiT)[37] for object detection [5, 22, 34], image segmentation [35, 54], video recognition [25]. Except for semantic segmentation with ViT[35, 47], most of the studies are motivated by the combined strength of UNet and ViT in medical image segmentation. The first ViT-based medical image segmentation model is TransUnet[6], that ViT encoders tokenized image patches from CNN, and decoders upsample the encoded features. SwinUNet[4] further replace the decoder with Swin-Transformer[24] blocks as a pure ViT-based encoder-decoder segmentation network. The most latest work [48] demonstrates to the strength of CNN and ViT can be simultaneously achieved in intra-slice and inter-slice of CT images. For a fair comparison of CNN and ViT in this SSL study, we propose to use a simple U-shape Encoder-Decoder segmentation backbone network with two network blocks i.e. pure CNN- or ViT-based block[4, 11].

2.2 Semi-Supervised Segmentation

The goal of Semi-Supervised Segmentation is to utilize the finite number of labelled images and high proportion of raw images to train the segmentation model. **Consistency regularization** of SSL has been widely studied which enforces the consistency of inference with various perturbations[13, 36, 39, 50]. A part of perturbation studies are applied to input images augmenting the images randomly and setting the consistency constraint among inferences of augmented images[13, 20]. The perturbation can also be applied to feature information. For example, the feature perturbation scheme is to develop multiple decoders and the difference of inferences of multi-decoder is trained to be similar to cross consistency training[30]. The most common SSL of consistency-aware with perturbation scheme is Student-Teacher style framework, which consists of a student model, and a teacher model [21, 36, 42]. The student is initialized with labelled data with perturbations, and the parameters of the teacher are updated from the student with averaging model weight. The teacher, hereby, is more robust than student, and can guide the learning of students with pseudo labels under the consistency-aware concern [50]. Except for the consistency regularization that aims to minimize the differences of images with perturbations, **adversarial learning**[16, 28] normally utilized an additional discriminator model to extract statistical features aiming to distinguish ground truth segmentation and inferences of segmentation model. In an adversarial learning process, the segmentation model and the discriminator model are trained against each other separately and iteratively. The discriminator aims to distinguish the high-quality segmentation inferences as the pseudo label, and the segmentation model aims to infer confidently from raw images against discriminator [16, 29, 52]. Motivated by the recent success of ViT, only a few studies report to train ViT in the SSL fashion for medical image segmentation starting from 2021. [27] proposed to combine the strength of ViT and CNN via co-training allowing ViT and CNN to complement and benefit each other with pseudo labels. [42] proposed to explore the ViT with a Student-Teacher style SSL, and an uncertainty-aware scheme is introduced when the teacher guides the student. In this paper, we further explore the ViT in SSL with an additional discriminator for adversarial training and a MixUp-based approach for interpolation consistency training to boost the performance of sViT.

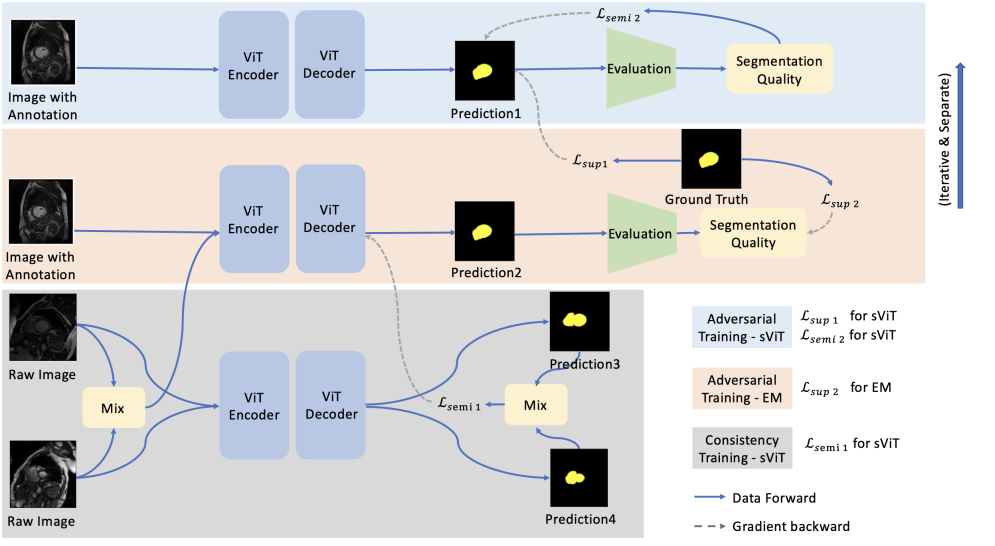

Figure 1: MixUp-based Interpolation Consistency & Adversarial Vision Transformer for Semi-Supervised Medical Image Semantic Segmentation. The training of proposed CAA-ViT consists of three stages including interpolation consistency training for sViT, two adversarial training stages for sViT and EM, respectively. Two supervision loss, and two semi-supervision loss are indicated accordingly.

3 Approaches

In a general medical image segmentation with the SSL fashion study, $\mathbf{L}$, $\mathbf{U}$ and $\mathbf{T}$ normally denote labelled training data set, unlabelled training data set, and a testing data set. We denote a batch of labelled 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 labelled training set and labelled testing set with its corresponding ground truth, and a batch of only raw data as $(\boldsymbol{X}_{\mathrm{u}}) \in \mathbf{U}$ in the unlabelled training set, where $\boldsymbol{X} \in \mathbb{R}^{h \times w}$ representing a 2D gray-scale image. $\boldsymbol{Y}_{\mathrm{p}}$ is the dense map predicted by a segmentation model $f(\theta) : \boldsymbol{X} \mapsto \boldsymbol{Y}_{\mathrm{p}}$ with the θ as the parameters of the model f. $\boldsymbol{Y}_{\mathrm{p}}$ can be considered as a batch of pseudo label for unlabelled data $(\boldsymbol{X}_{\mathrm{u}}, \boldsymbol{Y}_{\mathrm{p}}) \in \mathbf{U}$ to retrain models. Final evaluation results are calculated based on the differences between Y_{p} and Y_{gt} of $\mathbf{T}$. The training objective of CAA-ViT is to minimize the supervision loss $Loss_{\mathrm{sup}}$ and the semi-supervision loss $Loss_{\mathrm{semi}}$ of sViT, and supervision loss $Loss_{\mathrm{sup}}$ of EM. The loss $Loss_{\mathrm{sup}}$ and $Loss_{\mathrm{semi}}$ differs from each other by using labelled or unlabelled data $\boldsymbol{X}_{\mathrm{l}}$ or $\boldsymbol{X}_{\mathrm{u}}$. The framework of CAA-ViT including adversarial training stage and MixUp-based interpolation consistency training stage with corresponding *Loss* are briefly illustrated in Figure 1. The CAA-ViT is mainly motivated by adversarial training[52], and consistency training[36, 39, 44, 45] to explore the power of vision transformer. Three different training stages including one MixUp-based interpolation consistency training stage for sViT, and two adversarial training stages separately for sViT and EM to against each other in an iterative manner are detailed discussed Section 3.1, 3.2, and 3.3.

3.1 Adversarial Training

In the adversarial training stage, two models are introduced i.e. a ViT-based segmentation model(sViT) $f_{ViT}(\theta) : \boldsymbol{X} \mapsto \boldsymbol{Y}_{\mathrm{p}}$, and a CNN-based evaluation model(EM) $f_{CNN}(\theta) : f_{ViT}(\boldsymbol{X};\theta) \mapsto \boldsymbol{Y}_{\mathrm{quality}}$ to evaluate the quality of inference. The ViT-based model takes the input images $\boldsymbol{X}$ and infers corresponding segmentation feature maps $\boldsymbol{Y}_{\mathrm{p}}$, while incorporating suggestive loss from the EM, also known as the discriminator [16, 29, 52]. The EM takes the inference of sViT $\boldsymbol{Y}_{\mathrm{p}}$, and the corresponding input images $\boldsymbol{X}$ and infers a quality score $\boldsymbol{Y}_{\mathrm{quality}}$. In the CAA-ViT, the discriminator is designed to classify the quality of $\boldsymbol{Y}_{\mathrm{p}}$ with $\boldsymbol{Y}_{\mathrm{quality}} \in (0,1)$, which is considered as a binary classification task, where 1: high-quality inference, images with annotations, 0: low-quality inference, images without annotations. The architecture of EM is with VGG16 [33].

In either of the two adversarial training stages of sViT and EM as shown in Figure 1, one model is trained and the other model is evaluated. The goal of the EM is to distinguish the inference of sViT from using annotated input or unannotated input[52]. The sViT, however, is encouraged to make predictions from unannotated input with the same quality as from annotated input. In the training process, the sViT is trained by taking judgement from EM to improve the quality of inference achieving the same quality regardless of unannotated, while the EM is trained to discern the difference between the results from annotated input and the results from unannotated input since indices of input with labels are known. By doing so, both two models, hereby, achieve good segmentation and discriminating ability after training. The details of adversarial training objectives(loss of each model) is discussed in Section 3.3

3.2 MixUp-Based Interpolation Consistency Training

In the consistency training stage of sViT, we introduce a MixUp-based interpolation consistency training approach. Motivated by consistency regularization, which is to make the model robust against various perturbations on the unlabelled data [2, 29, 31], pixel-wise perturbation is introduced for sViT consistency training. Following the main idea of [39] to regularizes SSL by encouraging consistent predictions shown in Equation 1,

$$f(\alpha X_{u1} + (1-\alpha)X_{u2}) = \alpha f(X_{u1}) + (1-\alpha) f(X_{u2}) \tag{1}$$

where X_{u1}, X_{u2} are two randomly selected raw images as the input of model f, because two samples are likely to help find the decision boundary between the two classes with numerous and balanced classes [39]. α is a hyperparameter. Given MixUp operation [51] which is illustrated in Equation 2,

$$Mix_{\lambda}(a,b) = \lambda \times a + (1-\lambda) \times b \tag{2}$$

and pixel-wise interpolation, the sViT f_{ViT} is enforced to provide consistent predictions at interpolations of unlabelled images. The objective of consistency training stage aims to ensure reliable and consistency segmentation of image points interpolated from existing points, hereby, can be summarized as Equation 3

$$f_{ViT}(Mix_{\lambda}(X_{u1}, X_{u2}); \theta) \approx Mix_{\lambda}(f_{ViT}(X_{u1}; \overline{\theta}), f_{ViT}(X_{u2}; \overline{\theta})) \tag{3}$$

where the MixUp-based consistency is maintained between the outputs of the interpolated inputs $f_{ViT}(Mix_{\lambda}(X_{u1}, X_{u2}); \theta)$, and the interpolated outputs of the original inputs

$Mix_\lambda(f_{ViT}(X_{u1};\overline{\theta}), f_{ViT}(X_{u2};\overline{\theta}))$. $\overline{\theta}$ is the exponential moving average of θ [36]. The details of consistency training objectives is discussed in Section 3.3

3.3 Objective

The training objective is to minimize the sum of the supervision loss $\mathcal{L}_{\text{sup}}$ and the semi-supervision $\mathcal{L}_{\text{semi}}$ of three proposed training stages as shown in Equation 4

$$\mathcal{L} = \mathcal{L}_{\text{sup1}} + \mathcal{L}_{\text{sup2}} + \lambda(\mathcal{L}_{\text{semi1}} + \mathcal{L}_{\text{semi2}}) \tag{4}$$

where $\mathcal{L}_{\text{sup}}$ or $\mathcal{L}_{\text{semi}}$ depends on whether the data is annotated. λ is with ramp-up function [21], as $\lambda = e^{-5\times(1-t_{\text{iteration}}/t_{\text{maxiteration}})^2}$, where t is the iteration. λ is updated every 150 iterations, which is to make the training move the focus from $\mathcal{L}_{\text{sup}}$ to initialize mode with annotations to $\mathcal{L}_{\text{semi}}$ to learn from features of raw images. The $\mathcal{L}_{\text{sup1}}$ is to train the sViT with labelled data $(\boldsymbol{X}_{\text{l}}, \boldsymbol{Y}_{\text{gt}}) \in \mathbf{L}$ in one of the adversarial training stage, illustrated in Equation 5

$$\mathcal{L}_{\text{sup1}} = \text{CE}(Y_{\text{gt}}, f_{ViT}(\boldsymbol{X}_l;\theta)) + \text{Dice}(Y_{\text{gt}}, f_{ViT}(\boldsymbol{X}_l;\theta)) \tag{5}$$

where *CE* and *Dice* demonstrates the CrossEntropy-based and DiceCoefficient-based difference measures. The $\mathcal{L}_{\text{sup2}}$ is to train the EM with labelled data $(\boldsymbol{X}_{\text{l}}, \boldsymbol{Y}_{\text{gt}}) \in \mathbf{L}$, unlabelled data $(\boldsymbol{X}_{\text{u}}) \in \mathbf{U}$, and the inference of sViT in one of the adversarial training stage, illustrated in Equation 6

$$\mathcal{L}_{\text{sup2}} = \text{BCE}(f_{CNN}(f_{ViT}(\boldsymbol{X}_l/\boldsymbol{X}_u;\theta_{ViT});\theta_{CNN}), 1/0)) \tag{6}$$

here, Binary Cross-Entropy(*BCE*) is selected considering the EM is to classify the quality of inference of sViT. The score 1 indicates the inference is with high quality or from an annotated image, and score 0 indicates the inference is with low quality or from an unannotated image. The $\mathcal{L}_{\text{semi1}}$ is to train the sViT with unlabelled data $(\boldsymbol{X}_{\text{u}}) \in \mathbf{U}$ in the consistency training stage, illustrated in Equation 7

$$\mathcal{L}_{\text{semi1}} = \text{CE}(f_{ViT}(Mix_\lambda(X_{u1}, X_{u2});\theta), f_{ViT}(\boldsymbol{X}_l;\theta), Mix_\lambda(f_{ViT}(X_{u1};\overline{\theta}), f_{ViT}(X_{u2};\overline{\theta}))) \tag{7}$$

where CE is solely utilized on unlabelled data, such as X_{u1}, and X_{u2}. In light of the inspiration of using MixUp in supervised setting to gain large decision boundary margin [51], here in unsupervised setting MixUp-based interpolation is applied to sample X_{u1}, and X_{u2}, and corresponding pseudo labels $\boldsymbol{Y}_{\text{p}} = f_{ViT}(X_u;\overline{\theta})$. Thus the parameter θ is moving to make inference of sViT closer to the pseudo label. $\mathcal{L}_{\text{semi1}}$ is to train the sViT with the quality score of EM in one of the consistency training stage, illustrated in Equation 8

$$\mathcal{L}_{\text{semi2}} = -\text{CE}(f_{CNN}(f_{ViT}(\boldsymbol{X}_u;\theta_{ViT}), \boldsymbol{X}_u;\theta_{CNN}), 0) \tag{8}$$

By applying *CE* for evaluation score of inference $f_{ViT}(\boldsymbol{X}_u)$ from EM f_{CNN}, the sViT is enforced to predict reasonably good segmentation maps supervised by EM.

4 Experiments

Data sets: Our experiments validate the CAA-ViT and baseline methods on a public available MRI cardiac segmentation set, also known as ACDC, from MICCAI Challenge 2017 [1]. The data is collected from 100 patients with multi-clasee annotation including right ventricle, the endocardial and epicardial walls of the left ventricle covering different distributions of feature information, across five evenly distributed subgroups: normal, myocardial infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy, and abnormal right ventricle. All images are resized to 224×224 following the requirement of input of ViT segmentation backbone network[4]. 20% of images are selected as testing set, and the rest of data set is for training(including validation). There is no overlap between test, labelled train, unlabelled train, and validation set.

Implementation Details: Our code has been developed under Ubuntu 20.04 in Python 3.8.8 using Pytorch 1.10 and CUDA 11.3 using one Nvidia GeForce RTX 3090 GPU, and Intel(R) Intel Core i9-10900K. The runtimes averaged around 3.5 hours, including the data transfer, training, and inference. Data set is processed for 2D image segmentation purpose. CAA-ViT is trained for 30,000 iterations, batch size is set to 24, optimizer is SGD[3], and learning rate is set to 0.01, momentum is 0.9, and weight decay is 0.0001. The network is evaluated on validation set every 200 iterations, and the weight of network is saved only when the performance on validation outperforms the previous best performance. The above setting is also applied to all other baseline methods directly without any modification.

Backbone: The CAA-ViT consists of two models as shown in Figure 1. One is Swin-ViT-based U-shape encoder-decoder style segmentation model with skip connection[11, 24], and the other one is CNN-based discriminator following [33]. For a fair comparison, ViT-based U-shaped model selected in this study is aiming to make the computational cost and training efficiency similar with CNN-based UNet. The additional results of SSL methods with UNet results in CAA-CNN and other baseline methods are reported in Appendix.

Baseline Methods: All methods including CAA-ViT and other baseline methods are trained with the same hyper-parameter setting and the same distribution of features. The random selection of test set, labelled train set and unlabelled train set was only conducted once and then tested with all baseline methods together as well as CAA-ViT. The baseline methods reported includes: MT [36], DAN [52], ICT [39], ADVENT [40], UAMT [50], DCN [32] with ViT[4] and CNN[11] as the segmentation backbone network.

Evaluation Metrics: The direct comparison experiments between CAA-ViT and other baseline methods are conducted with a comprehensive f evaluation metrics including similarity measures(the higher the better): Dice, IOU, Accuracy, Precision, Sensitivity, Specificity, and difference measures(the lower the better): Hausdorff Distance (HD) 95%, Average Surface Distance (ASD). The mean value of these metrics is reported because the data set is a multi-class segmentation data set. The full evaluation measures are reported when comparing CAA-ViT against other baseline methods. IOU as the most common metric is also selected to report the performance of all baseline methods and CAA-ViT under different assumptions of the ratio of labelled/total data including 10%, 20%, 30%, 60%, 70%, 80%, 90%, 100%(fully supervised), and used to report the evaluation of each inference on the test set with a distribution chart.

Quantitative Results: Table 1 reports the direct comparison of CAA-ViT against other SSL methods including similarity measures and difference measures when the ratio of assumed labelled/total data is 10%. Table 2 reports the CAA-ViT and other SSL methods performance under different assumptions of ratio of labelled/total data. The best result on table 1,

2 is in **Bold**. A line chart in logarithmic scale is briefly sketched in Figure 3 (a), where X-axis is the ratio of labelled/total data, and Y-axis is the IOU performance, illustrating the valuable performance of CAA-ViT against other baseline methods especially in low ratio of labelled/total data. IOU is selected to evaluate each image of test set, and a histogram indicating the cumulative distribution of IOU performance of inference is briefly sketched in Figure 3 (b), where the X-axis is the IOU threshold and the Y-axis is the number of predicted images on test set, demonstrating CAA-ViT more likely to predict images with high IOU against other methods. More details of the qualitative results are reported in Appendix.

Qualitative Results: Figure 2 illustrates eight randomly selected sample raw images, corresponding inference against the publish ground truth of all baseline methods including CAA-ViT, where Yellow, Red, Green and Black denote True Positive(TP), False Positive(FP), False Negative(FN) and True Negative(TN) region on test set at pixel level.

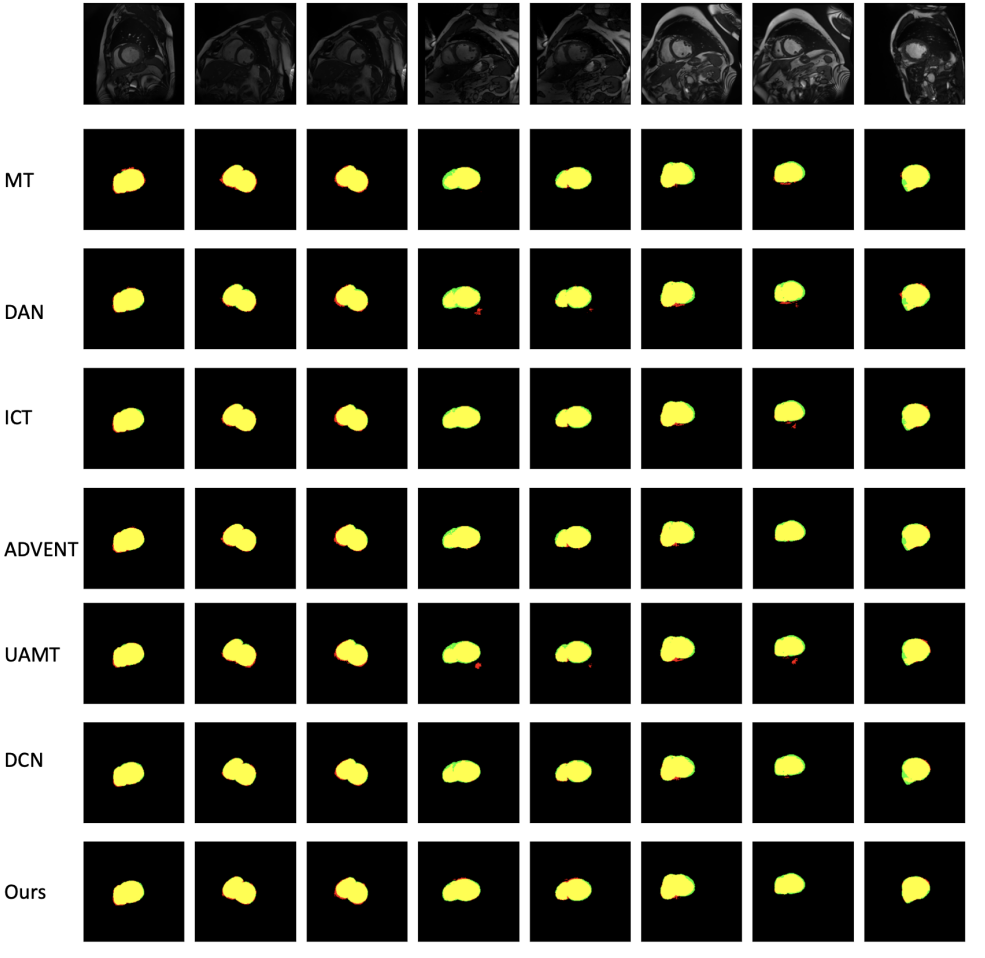

Figure 2: The Qualitative Inference Results of CAA-ViT and Other Methods Against Ground Truth Under 10% of Label/Total Data for Training

Table 1: Direct Comparison of Semi-supervised Frameworks on MRI Cardiac Test Set

Framework	mDice↑	mIOU↑	Acc↑	Pre↑	Sen↑	Spe↑	HD↓	ASD↓
MT[36]+ViT	0.8384	0.7359	0.9939	0.8362	0.8415	0.9716	23.7689	2.2801
DAN[52]+ViT	0.8232	0.7165	0.9932	0.8243	0.8222	0.9686	23.5824	2.8411
ICT[39] +ViT	0.8663	0.7748	0.9948	0.8607	0.8722	**0.9775**	22.5653	2.2404
ADVENT[40]+ViT	0.8654	0.7733	0.9949	0.8689	0.8625	0.9743	20.4493	1.9022
UAMT[50] +ViT[42]	0.8542	0.7575	0.9945	0.8523	0.8575	0.9752	23.6615	2.2159
DCN[32] +ViT	0.8728	0.7844	0.9950	0.8770	0.8690	0.9738	22.3759	1.9762
CAA-ViT(ours)	**0.8824**	**0.7984**	**0.9954**	**0.8887**	**0.8765**	0.9748	**18.9200**	**1.7587**

Table 2: The Mean IOU Results on Test Set Under Different Assumptions of Ratio of Label/Total Data for Training

labelled/Total	10%	20%	30%	60%	70%	80%	90%	100%
MT[36]+ ViT	0.7359	0.8096	0.8338	0.8567	0.8569	0.8593	0.8596	0.8655
DAN[52]+ ViT	0.7165	0.7927	0.8148	0.8451	0.8467	0.8432	0.8391	0.8544
ICT[39] + ViT	0.7748	0.8147	0.8318	0.8556	0.8583	0.8545	0.8637	0.8658
ADVENT[40]+ ViT	0.7733	0.8099	0.8315	0.8581	0.8610	0.8561	0.8577	0.8682
UAMT[50]+ ViT[42]	0.7575	0.8085	0.8291	0.8541	0.8538	0.8571	0.8614	0.8672
DCN[32] + ViT	0.7844	**0.8248**	0.8302	0.8604	0.8565	0.8582	0.8594	0.8665
CAA-ViT(ours)	**0.7984**	0.8192	**0.8485**	**0.8681**	**0.8670**	**0.8674**	**0.8748**	**0.8775**

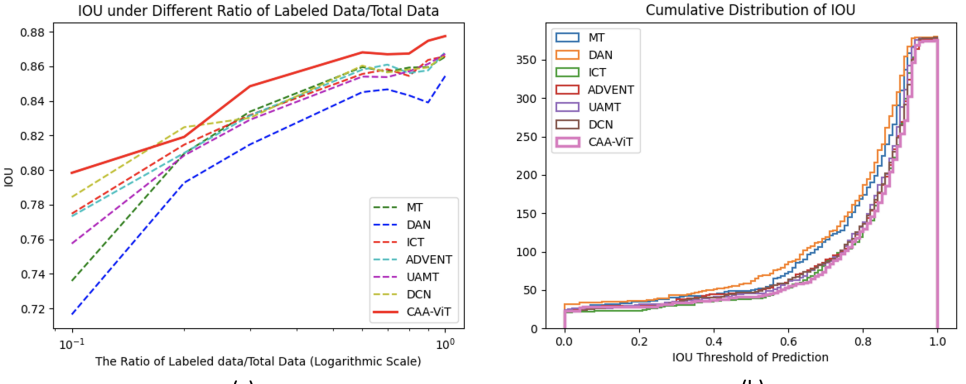

Figure 3: The Performance of Proposed Model with Baseline Methods for Medical Image Segmentation

5 Conclusion

In this paper, the exploration of a segmentation vision transformer in an advanced semi-supervision fashion, CAA-ViT, is introduced. A novel adversarial training and a MixUp-based interpolation consistency training are designed simultaneously that allow vision transformer to benefit from a large amount of medical data with limited annotations. In terms of performance, CAA-ViT outperforms other methods with a variety of measures under different situations of ratio labelled/total data with promising results.

References

[1] Olivier Bernard 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.

[2] Gerda Bortsova, Florian Dubost, Laurens Hogeweg, Ioannis Katramados, and Marleen de Bruijne. Semi-supervised medical image segmentation via learning consistency under transformations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 810–818. Springer, 2019.

[3] Léon Bottou. Stochastic gradient learning in neural networks. In *Proceedings of Neuro-Nîmes 91*, Nimes, France, 1991. EC2.

[4] Hu Cao et al. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[6] Jieneng Chen et al. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] K He et al. Deep residual learning for image recognition. In *Proc IEEE CVPR*, pages 770–778, 2016.

[10] O Oktay et al. Attention U-Net: Learning where to look for the pancreas. *Int Conf Medical Imaging with Deep Learning*, 2018.

[11] O Ronneberger et al. U-Net: Convolutional networks for biomedical image segmentation. In *Int Conf Med Im Comp & Comp-Assisted Intervention*, pages 234–241. Springer, 2015.

[12] S Woo et al. CBAM: Convolutional block attention module. In *Proc ECCV*, pages 3–19, 2018.

[13] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019.

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[15] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[16] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*, 2018.

[17] Nabil Ibtehaz and M Sohel Rahman. Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural Networks*, 121:74–87, 2020.

[18] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[19] Fabian Isensee et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. In *Bildverarbeitung für die Medizin 2019*, pages 22–22. Springer, 2019.

[20] Jongmok Kim, Jooyoung Jang, and Hyunwoo Park. Structured consistency loss for semi-supervised semantic segmentation. *arXiv preprint arXiv:2001.04647*, 2020.

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

[22] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022.

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proc IEEE CVPR*, pages 2117–2125, 2017.

[24] Ze Liu, Yutong Lin, et al. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[25] Ze Liu et al. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022.

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

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

[28] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE transactions on pattern analysis and machine intelligence*, 43(4):1369–1379, 2019.

[29] Takeru Miyato et al. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

[30] 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.

[31] Sungrae Park, JunKeon Park, Su-Jin Shin, and Il-Chul Moon. Adversarial dropout for supervised and semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[32] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *Proceedings of the European Conference on Computer Vision*, pages 135–152, 2018.

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Int Conf Learning Representations(ICLR)*, 2015.

[34] Hwanjun Song et al. Vidt: An efficient and effective fully transformer-based object detector. *arXiv preprint arXiv:2110.03921*, 2021.

[35] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. *arXiv preprint arXiv:2105.05633*, 2021.

[36] Antti Tarvainen and Harri Valpola. 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*, pages 1195–1204, 2017.

[37] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[39] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *International Joint Conference on Artificial Intelligence*, pages 3635–3641, 2019.

[40] 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.

[41] Ziyang Wang. Computationally-efficient vision transformer for medical image semantic segmentation via dual pseudo-label supervision. In *IEEE International Conference on Image Processing (ICIP)*, 2022.

[42] Ziyang Wang. An uncertainty-aware transformer for mri cardiac semantic segmentation via mean teachers. *Annual Conference on Medical Image Understanding and Analysis*, 2022.

[43] Ziyang Wang and Irina Voiculescu. 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.

[44] Ziyang Wang and Irina Voiculescu. Triple-view feature learning for medical image segmentation. In *MICCAI Workshop on Resource-Efficient Medical Image Analysis*, pages 42–54. Springer, 2022.

[45] Ziyang Wang, Tianze Li, Jian-Qing Zheng, and Baoru Huang. When cnn meet with vit: Towards semi-supervised learning for multi-class medical image semantic segmentation. *arXiv preprint arXiv:2208.06449*, 2022.

[46] Ziyang Wang et al. 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.

[47] Enze Xie et al. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[48] Xiangyi Yan et al. After-unet: Axial fusion transformer unet for medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3971–3981, 2022.

[49] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[50] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–613. Springer, 2019.

[51] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[52] Yizhe Zhang, Lin Yang, Jianxu Chen, Maridel Fredericksen, David P Hughes, and Danny Z Chen. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *International conference on medical image computing and computer-assisted intervention*, pages 408–416. Springer, 2017.

[53] Hengshuang Zhao et al. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[54] Sixiao Zheng et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.