6 Appendix / Supplementary Material

6.1 More Details of Experiments and Code

The CNN- and ViT-based segmentation backbone networks are utilized to fairly explore our SSL methods. The CNN, ViT networks both are with the same U-Shape style architectures based on pure CNN or ViT blocks, which results in conventional U-Net, and Swin-Transformer-based U-Net. There is no further modifications of the segmentation backbone, and the computational cost of the two segmentation models are similar. All the SSL baseline methods and proposed CAA-ViT are developed and tested with the same hyperparameter setting including optimizer, learning rate, batch size, loss function and etc. The feature distribution of labelled data set, unlabelled data set, validation data set, and test data set is the same for all methods in experiment section(i.e. the randomly selections of labelled train set, unlabelled train set, validation set, test set from the public data set are only conducted once). The code will be public available.

6.2 More Details of Qualitative Results

Figure 4 illustrated in the same way following in Figure 2 but sketch CAA-ViT under the different assumption of ratio of labelled/total data for training(quantitative results are reported in Table 2), demonstrating the CAA-ViT is able to predict less FP, and FN pixels of test set when more annotations provided.

6.3 More Details of Quantitative Results

What if we directly test conventional CNN with proposed SSL methods? This paper explore the power of ViT for SSL segmentation. Following the exploration of ViT in SSL fashion in Table 1 and Table 2, we further demonstrate our proposed CAA-ViT can be dominated in CNN-based(UNet) SSL methods as well, and reported in Table 3, and Table 4 as CAA-CNN. All images are resized to 256×256 for training and testing the CNN-based model.

What if the ratio of labelled/total data is extremely low? Table 1 reports the CAA-ViT and other baseline methods under the assumption of 10% train set is labelled. We also test the extremely low ratio of labelled/total data situation, i.e. 1%, in Table 5. It is valuable to see that CAA-ViT outperforms other methods under 1% of labelled/total data situation, and demonstrating the efficiency of proposed contribution of SSL including: adversarial training, and consistency-aware training.

What if we evaluate each inference of all methods on the test set? In Table 1, the mean value of evaluation measures on test set is reported. Each of image on test set has also been evaluated with IOU, and the distribution of IOU is reported in Figure 5, where CAA-ViT demonstrates more likely to predict with high IOU inference.

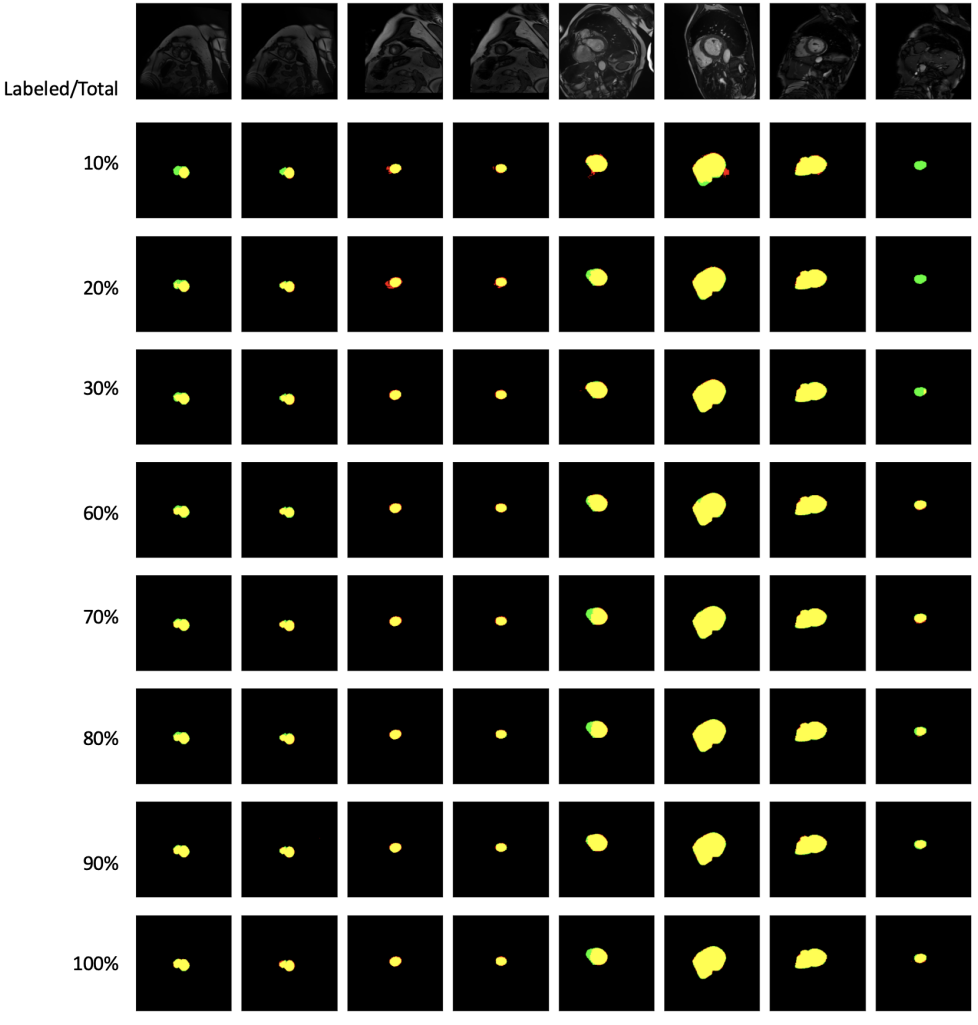

Figure 4: The Qualitative Inference Results of CAA-ViT Against Ground Truth Under Different Assumptions of Ratio of Label/Total Data for Training

Table 3: Direct Comparison of Semi-supervised Frameworks with CNN on MRI Cardiac Test Set

Framework	mDice↑	mIOU↑	Acc↑	Pre↑	Sen↑	Spe↑	HD↓	ASD↓
MT[36]+CNN	0.8860	0.8034	0.9952	0.8898	0.8829	0.9720	9.3659	2.5960
DAN[52]+CNN	0.8773	0.7906	0.9947	0.8721	0.8832	0.9743	9.3203	3.0326
ICT[39] +CNN	0.8902	0.8096	0.9954	0.8916	0.8897	0.9745	11.6224	3.0885
ADVENT[40]+CNN	0.8728	0.7836	0.9947	0.8985	0.8517	0.9601	9.3203	3.5026
UAMT[50] +CNN	0.8683	0.7770	0.9946	**0.8988**	0.8416	0.9582	8.3944	**2.2659**
DCN[32] +CNN	0.8809	0.7953	0.9951	0.8915	0.8714	0.9690	**8.9155**	2.7179
CAA-CNN(ours)	**0.8970**	**0.8197**	**0.9957**	0.8962	**0.8980**	**0.9765**	30.7428	2.3448

Table 4: The Mean IOU Results of CNN-based SSL Methods 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]+CNN	0.8034	0.8294	0.8397	0.8580	0.8680	0.8801	0.8700	0.8583
DAN[52]+CNN	0.7906	0.8130	0.8356	0.8596	0.8668	0.8683	0.8612	0.8780
ICT[39] +CNN	0.8096	0.8191	0.8512	0.8621	0.8601	0.8651	0.8671	0.8853
ADVENT[40] +CNN	0.7836	0.8133	0.8537	0.8595	0.8650	0.8644	0.8769	0.8797
UAMT[50] +CNN	0.7770	0.8269	0.8416	0.8655	0.8647	0.8605	0.8493	0.8778
DCN[32] +CNN	0.7953	0.8252	0.8455	0.8606	0.8637	0.8670	0.8711	0.8769
CAA-CNN(ours)	**0.8197**	**0.8430**	**0.8622**	**0.8785**	**0.8706**	**0.8806**	**0.8807**	**0.8810**

Table 5: Direct Comparison of Semi-supervised Frameworks on MRI Cardiac Test Set(only 1% of train set is labelled)

Framework	mDice↑	mIOU↑	Acc↑	Pre↑	Sen↑	Spe↑	HD↓	ASD↓
MT[36]+ViT	0.6108	0.4873	0.9852	0.6695	0.5795	0.9014	46.4367	20.3735
DAN[52]+ViT	0.6460	0.5198	**0.9853**	0.6531	0.6458	0.9258	59.3747	10.1346
ICT[39] +ViT	0.6368	0.5115	0.9852	0.6695	0.6218	0.9167	38.1371	7.7604
ADVENT[40]+ViT	0.6178	0.4937	0.9851	0.6530	0.5978	0.9113	**33.1885**	7.2299
UAMT[50] +ViT[42]	0.6158	0.4927	0.9848	0.6629	0.5954	0.9104	39.9434	8.6557
DCN[32] +ViT	0.6380	0.5126	0.9866	**0.7252**	0.6006	0.9092	38.6230	**3.8966**
CAA-ViT(ours)	**0.6992**	**0.5747**	0.9852	0.6735	**0.7350**	**0.9421**	63.5797	9.9284

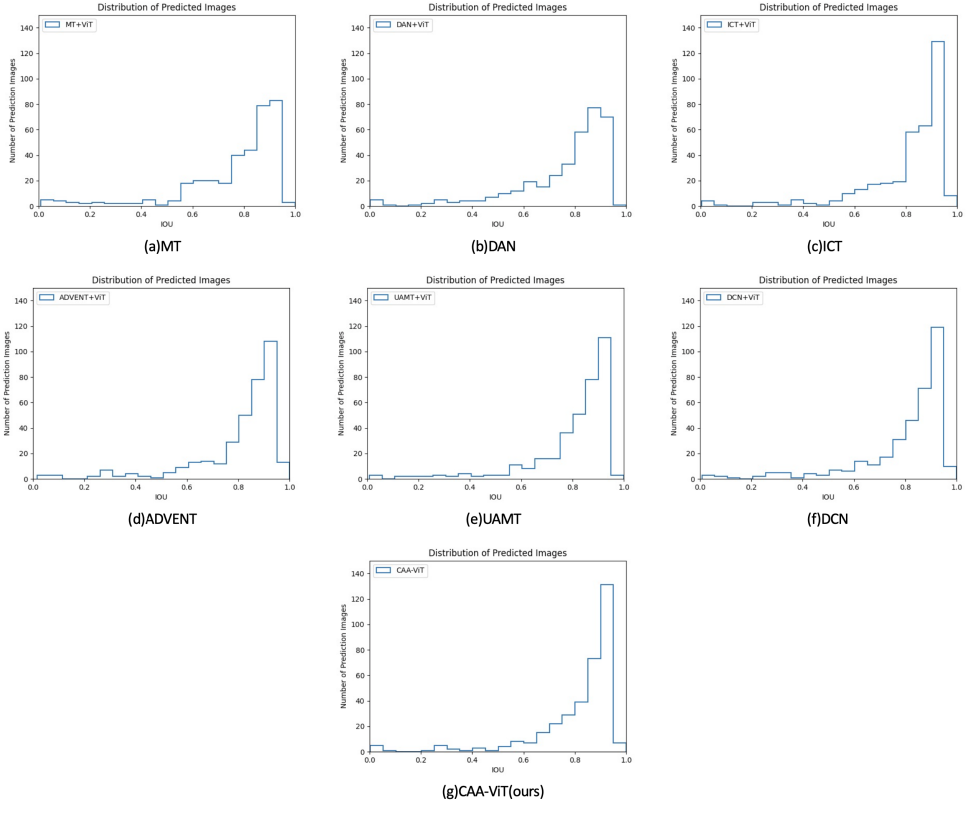

Figure 5: The IOU Distribution of Each Inference on Test Set