When CNN Meet with ViT: Towards Semi-Supervised Learning for Multi-Class Medical Image Semantic Segmentation

Ziyang Wang[1], Tianze Li[2], Jian-Qing Zheng[3], and Baoru Huang[4]

[1] Department of Computer Science, University of Oxford, UK
ziyang.wang@cs.ox.ac.uk
[2] Canford School, UK
[3] The Kennedy Institute of Rheumatology, University of Oxford, UK
[4] Department of Surgery and Cancer, Imperial College London, UK

Abstract. Due to the lack of quality annotation in medical imaging community, semi-supervised learning methods are highly valued in image semantic segmentation tasks. In this paper, an advanced consistency-aware pseudo-label-based self-ensembling approach is presented to fully utilize the power of Vision Transformer(ViT) and Convolutional Neural Network(CNN) in semi-supervised learning. Our proposed framework consists of a feature-learning module which is enhanced by ViT and CNN mutually, and a guidance module which is robust for consistency-aware purposes. The pseudo labels are inferred and utilized recurrently and separately by views of CNN and ViT in the feature-learning module to expand the data set and are beneficial to each other. Meanwhile, a perturbation scheme is designed for the feature-learning module, and averaging network weight is utilized to develop the guidance module. By doing so, the framework combines the feature-learning strength of CNN and ViT, strengthens the performance via dual-view co-training, and enables consistency-aware supervision in a semi-supervised manner. A topological exploration of all alternative supervision modes with CNN and ViT are detailed validated, demonstrating the most promising performance and specific setting of our method on semi-supervised medical image segmentation tasks. Experimental results show that the proposed method achieves state-of-the-art performance on a public benchmark data set with a variety of metrics. The code is publicly available.[1]

1 Introduction

Medical image segmentation is an essential task in computer vision and medical image analysis community where deep learning methods have shown dominated position recently. The promising results of current deep learning study not only relies on architecture engineering of CNN [28,34,9,43], but also on

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

sufficient high-quality annotation of data set [2,13,48,11]. The most common situation of clinical medical image data, however, is with a small amount of labelled data and a large number of raw images such as CT, ultrasound, MRI, and videos from laparoscopic surgery [30,17,44,47]. In recent studies of neural network architecture engineering, the performance of purely self-attention-based, Transformer [40], outperforms CNN and RNN because of the ability of modeling long-range dependencies [27,13]. Following the above concern of data situation, and the recent success in network architecture engineering, we hereby proposed a **S**emi-**S**upervised medical image **S**emantic **S**egmentation framework aiming to fully utilize the power of **C**NN and **V**iT simultaneously, called **S4CVnet**. The framework of S4CVnet consists of a feature-learning module, and a guidance module, which is briefly sketched in Figure 1. This setting is inspired by the Student-Teacher style framework [15,39,48], that the perturbation is applied to the student network, and the parameters of the teacher network are updated through Exponential Moving Average(EMA) [26] which makes the teacher network much robust to guide the learning of the student network with pseudo label under consistency-aware concern. To utilize the feature-learning power of CNN and ViT simultaneously and avoid the barrier caused by the different architecture of two networks, we hereby come up with a dual-view co-training approach in the feature-learning module[47,10]. Two different views of networks infer pseudo labels simultaneously to expand the size of the data set with raw data, complementing and beneficial to each other during the training process. One feature-learning network is also considered a bridge to be applied network perturbation and transfer of learning knowledge via the Student-Teacher style scheme.

The contributions of S4CVnet is fourfold and discussed as follows:

- an enhanced dual-view co-training module aiming to fully utilize the feature-learning power of CNN and ViT mutually is proposed. Both CNN and ViT are with the same U-shape Encoder-Decoder style segmentation network for fair comparison and exploration,
- a robust guidance module based on computational efficient U-shape ViT is proposed, and a consistency-aware Student-Teacher style approach via EMA is properly designed,
- an advanced semi-supervised multi-class medical image semantic segmentation framework is proposed, evaluated on a public benchmark data set with a variety of evaluation measures, and keeps state-of-the-art against other semi-supervised methods under the same setting and feature information distribution to our best of knowledge [39,52,41,42,51,45,32,30,47,48],
- a topological exploration study of all alternative supervision modes with CNN and ViT, as well as an ablation study, is validated to present a whole picture of utilizing CNN and ViT in a semi-supervised manner, and demonstrates the most proper setting and promising performance of S4CVnet.

2 Related Work

Semantic Segmentation The convolutional neural network(CNN) for image semantic segmentation, as a dense prediction task, has been widely studied since 2015, i.e. FCN [28]. It is the first CNN-based network trained with a supervised fashion for pixels-to-pixels prediction tasks. Then, the subsequent study of segmentation was dominated by CNN with three aspects of contribution: backbone network, network blocks, and training strategy. For example, one of the most promising backbone networks is UNet [34], which is an Encoder-Decoder style network with skip connections to efficiently transfer multi-scale semantic information. A variety of advanced network blocks to further improve CNN performance such as attention mechanism [35,24], residual learning [16], densely connected [18], dilated CNN [8] have been applied to the backbone network, UNet, which results in a family of UNet [21,44,46,23]. The CNN for dense prediction tasks, however, is lack of ability of modelling long-range dependencies in recent studies, and is defeated by Transformer, a purely self-attention-based network, that originated from natural language processing [40]. The Transformer was widely explored in computer vision tasks, i.e. Vision Transformer(ViT) [13], around classification, detection, and segmentation tasks [27,38,6,4]. In this paper on the backbone aspect, we focus on exploring the feature-learning power of CNN and ViT simultaneously, enabling both of them beneficial to each other, and specifically tackling a semi-supervised dense prediction task based on a multi-view co-training self-ensembling approach.

Semi-Supervised Semantic Segmentation Besides the study of backbone networks and network blocks, the training strategy is also an essential study depending on the different scenarios of data set, such as weakly-supervised learning to tackle low-quality annotations [53,5,36], noisy annotations [44], multi-rater annotations [25], and mixed-supervised learning for multi-quality annotations [33]. The most common situation of medical imaging data is with a small amount of labelled data and a large amount of raw data due to the high labelling cost, so semi-supervised learning is significantly valuable to be explored. Co-training, and self-training are two widely studied approaches in semi-supervised learning. Self-training, also known as self-labelling, is to initialize a segmentation network with labelled data at first. Then the pseudo segmentation masks on unlabelled data are generated by the segmentation network [50,7,20,54,31,29]. A condition is set for the selection of pseudo segmentation masks, and the segmentation network is retrained by expanding training data several times. GAN-based approaches mainly studied how to set the condition using discriminator learning for distinguishing the predictions and the ground-truth segmentation [19,37]. The other approach is Co-training, which is usually to train two separate networks as two views. These two networks thus expand the size of training data and complement each other. Deep Co-training was firstly proposed in [32] pointing out the challenge of utilizing co-training with a single data set, i.e. ‘collapsed neural networks’. Training two networks on the same data set cannot enable multi-view feature learning because two networks will necessarily end up similar. Disagreement-based Tri-net was proposed with three views which improved with diversity augmentation for pseudo label editing to solve ‘collapsed neural networks’ by ’View Differences’ [12,45]. Uncertainty estimation is also an approach to enable reliable pseudo labels to be utilized to train other views [49,48,10]. Current key studies of Co-training mainly on: (a)enabling the diversity of two views, and (b)properly/confidently generating pseudo labels for retraining networks. In this paper of the training strategy aspect, we adopt two completely different segmentation networks to encourage the difference between two views in the feature-learning module. Furthermore, inspired by the Student-Teacher style approach [39,26], a ViT-based guidance module is developed which is much more robust with the help of the feature-learning module via perturbation, and average model weights [26,14]. The guidance module is able to confidently and properly supervise the two networks of the feature-learning module in the whole semi-supervision process via pseudo label.

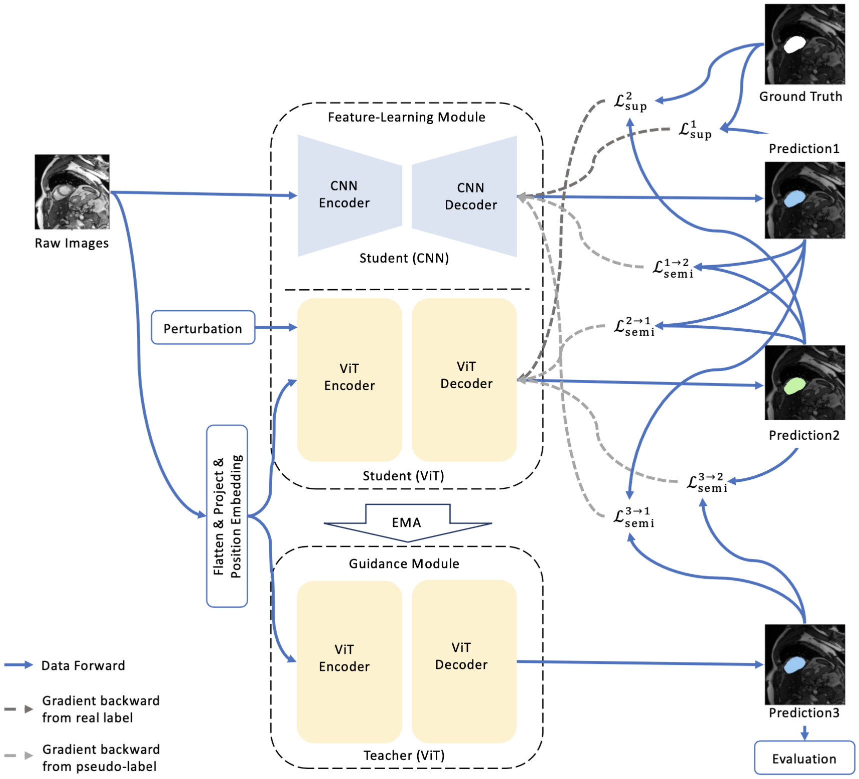


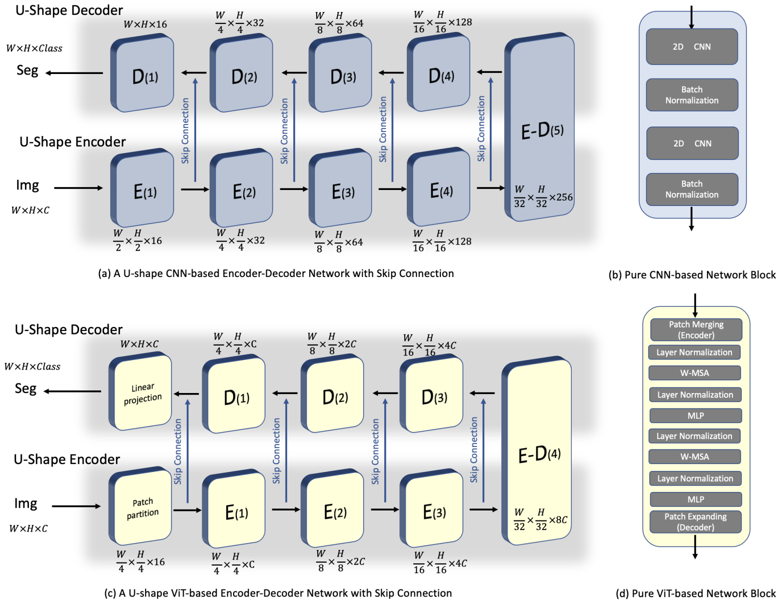
Fig. 1. The Framework of S4CVnet. It is a semi-supervised medical image semantic segmentation framework with the power of CNN and ViT, which consists of a feature-learning module(CNN & ViT), and a guidance module(ViT). The supervision mechanism is illustrated by minimizing the difference (also known as *Loss*) between prediction and (pseudo) labels.

3 Methodology

In generic semi-supervised learning for image segmentation tasks, $\mathbf{L}$, $\mathbf{U}$ and $\mathbf{T}$ normally denote a small number of labelled data, a large amount of unlabeled data, and a testing data set. 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 data 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 network $f(\theta) : \boldsymbol{X} \mapsto \boldsymbol{Y}_{\mathrm{p}}$ with the θ as the parameters of the network f. $\boldsymbol{Y}_{\mathrm{p}}$ can be considered as a batch of pseudo label for unlabeled data $(\boldsymbol{X}_{\mathrm{u}}, \boldsymbol{Y}_{\mathrm{p}}) \in \mathbf{U}$ for retraining networks. Final evaluation results are calculated based on the differences between Y_{p} and Y_{gt} of $\mathbf{T}$. The training of S4CVnet framework is to minimize the sum of supervision loss $Loss_{\mathrm{sup}}$ and the semi-supervision loss $Loss_{\mathrm{semi}}$ which are based on the difference of inference of each network with Y_{gt}, and Y_{p}, respectively. There is no overlap between $\mathbf{L}$, $\mathbf{U}$ and $\mathbf{T}$ in our study. The framework of S4CVnet, as shown in Figure 1, consists of a feature-learning module and a guidance module which are based on three networks f, i.e. a CNN-based network $f_{\mathrm{CNN}}(\theta)$, and two ViT-based networks $f_{\mathrm{ViT}}(\theta)$. The θ of each network of the feature-learning module are initialized separately to encourage the difference of the two views of learning, and the θ of the guidance module is updated from one of the feature-learning networks which have the same architecture via EMA. The final inference of S4CVnet is considered as the output by guidance module $f_{\mathrm{ViT}}(\overline{\theta}) : \boldsymbol{X} \mapsto \boldsymbol{Y}$. The details of CNN & ViT networks, feature-learning module, and guidance module are discussed in the following Section 3.1, 3.2, and Section 3.3, respectively.

3.1 CNN & ViT

To fairly compare, analyse, and explore the feature learning ability of CNN and ViT, we propose a U-shape encoder-decoder style multi-class medical image semantic segmentation network, and it can be built with a purely CNN-based network block or ViT-based network block, respectively. Motivated by the success of the skip connection of U-Net [34], we firstly propose a U-shape segmentation network with 4 encoders and decoders connected by skip connections which are briefly sketched in Figure 2 (a). A pure CNN or ViT segmentation network hereby can be directly built with replacing the encoders and decoders with the proposed network blocks which is sketched in Figure 2 (b). In each of CNN-based block, two 3×3 convolutional layers and two batch normalization [22] are developed accordingly [34]. The ViT-based block is based on Swin-Transformer block [27] with no further modification motivated by [6,4]. Different with the traditional Transformer block [13], layer normalization LN [1], multi-head self attention, residual connection [16], MLP with GELU are developed with shift-window which results in window-based multi-head self attention(WMSA) and shifted window-based multi-head self attention(SWMSA). Both of WMSA and SWMSA are applied in the two successive transformer blocks respectively shown



(a) A U-shape CNN-based Encoder-Decoder Network with Skip Connection

(b) Pure CNN-based Network Block

(c) A U-shape ViT-based Encoder-Decoder Network with Skip Connection

(d) Pure ViT-based Network Block

Fig. 2. The Backbone Segmentation Network. (a,c)a U-shape CNN-based or ViT-based encoder-decoder style segmentation network, (b,d)a pure CNN-based or ViT-based network block. These two network blocks can be directly applied to the U-shape encoder-decoder network resulting in a purely CNN- or ViT-based segmentation network.

on the upside of Figure 2 (b). The details of data pipeline through self-attention-based WMSA, SWMSA, MLP for feature learning of ViT are summarised in Equations 1, 2, 3, 4, and 5, where $i \in 1 \cdots L$, and L is the number of blocks. The self-attention mechanism comprises three point-wise linear layers mapping tokens to intermediate representations: quires $\boldsymbol{Q}$, keys $\boldsymbol{K}$, and values $\boldsymbol{V}$, introduced in Equation 5. In this way, the transformer block maps input sequence $\boldsymbol{Z}_0 = [z_{0,1} \cdots z_{0,N}]$ positions to $\boldsymbol{Z}_L = [z_{L,1}, ..., z_{L,N}]$, and the much richer sufficient semantic feature information(global dependencies) is fully extracted and collected through the ViT-based block.

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

$$\boldsymbol{Z}_{i} = \mathrm{MLP}(\mathrm{LN}(\boldsymbol{Z}_{i})) + \boldsymbol{Z}_{i} \tag{2}$$

$$\boldsymbol{Z}_{i+1} = \mathrm{SWMSA}(\mathrm{LN}(\boldsymbol{Z}_{i})) + \boldsymbol{Z}_{i} \tag{3}$$

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

$$\mathrm{MSA}(\boldsymbol{Z}') = \mathrm{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}}{\sqrt{D}})\boldsymbol{V} \tag{5}$$

where $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{M^2 \times d}$, and M^2 represents the number of patches in a window, and d is the dimension of the query and key.

Unlike conventional CNN-based blocks with downsampling and upsampling between each encoder or decoder, merging layers and expanding layers are designed between each ViT-based encoder, or decoder, respectively [6,4]. The merging layer is designed to reduce 2 times of the number of tokens and increase 2 times of the feature dimension. It divides the input patches into 4 parts and concatenates them together. A linear layer is applied to unify the dimension to 2 times. The expanding layer is designed to reshape the size of input feature maps 2 times bigger, and reduces the feature dimension to half of the input feature map dimension. It uses a linear layer to increase the feature dimension, and then rearranges operation is used to expand the size and reduce the feature dimension to a quarter of the input dimension. A brief illustration of the size of the feature map in each step is in Figure 2 (a), where W, H, C represents the width, height, and channel dimension of a feature map. Considering making the ViT the same computational efficiency with the CNN for a fair comparison and complement each other, we come up with the setting: patch size is 4, input channel is 3, embedded dimension is 96, the number of head of self-attention is 3,6,12,24, the window size is 7, 2 swin-transformer-based blocks for each encoder/decoder, and the ViT is pre-trained with ImageNet [11]. More details of CNN and ViT backbone with setting is available in Appendix.

3.2 Feature-Learning Module

The semi-supervised learning, especially in the pseudo-label-based approach, has been studied in image segmentation [3,32,12]. It incorporates segmentation inference on unlabeled data from one network as the pseudo label to retrain the other network, i.e. multi-view co-training approach [15,30]. Motivated by the recent success of cross pseudo label supervision [10], which has two networks $f(\theta_1)$, $f(\theta_2)$ with same architecture but initialized separately to encourage the difference of dual views, we further propose a feature-learning module aiming to explore the power of ViT and CNN mutually. Besides parameters of two networks are of course initialized separately, two completely different architectures of networks $f_{\mathrm{CNN}}(\theta_1)$, $f_{\mathrm{ViT}}(\theta_2)$ are designed to benefit each other via multi-view learning thus boost the performance of dual-view learning. The proposed feature-learning module to generate pseudo label can be illustrated as:

$$P_1 = f_{\mathrm{CNN}}(X; \theta_1), P_2 = f_{\mathrm{ViT}}(X; \theta_2). \tag{6}$$

where θ_1, θ_2 demonstrate network are initialized separately, P_1, P_2 represent the segmentation inference with $f_{\mathrm{CNN}}(\theta_1)$, $f_{\mathrm{ViT}}(\theta_2)$, respectively. The pseudo label based on P_1, P_2 then is utilized to supervise and complement each other. The CNN is mainly based on the local convolution operation, but the ViT is to model the global dependencies of feature through self-attention [13], so two segmentation inferences P_1, P_2 have different properties of prediction, and no explicit constraints to enforce two inferences similar. The supervision detail of simultaneously complementing each other (update parameters of ViT and CNN) is discussed in Section 3.4.

3.3 Guidance Module

Except for the feature-learning module to enable two networks to learn from the data, a robust guidance module is designed under the consistency-aware concern to boost the performance and also act as the final module for evaluation of S4CVnet. Inspired by temporal ensembling [26], and self-ensembling [39], the guidance network is to further supervise the perturbed networks and minimize the inconsistency. In a training process, the perturbation is firstly applied to the one of a network in the feature-learning module. Secondly, the parameter of the network is updated iteratively with back prorogation. Then the network of guidance module is updated via exponential moving average(EMA) from the feature-learning module. Finally, a much more robust guidance module which is more likely to be correct than the feature learning network is then to supervise two feature-learning networks with the consistency concern. In S4CVnet, the guidance module is based on ViT which has the same architecture as the ViT in the feature-learning network, so that guidance ViT can be constantly updated through EMA of the parameter of the ViT network learning from the data [26]. The proposed guidance module to generate pseudo label can be illustrated as:

$$P_3 = f_{\text{ViT}}(X; \overline{\theta}). \tag{7}$$

where $\overline{\theta}$ demonstrates the network ViT is based on averaging network weights rather than directly trained by the data. $\overline{\theta}$ is updated based on the parameter of feature-learning ViT model θ_t on past training step t, which can be illustrated as $\overline{\theta} = \alpha\theta_{t-1} + (1-\alpha)\theta_t$. α is a weight factor which is calculated as the $\alpha = 1 - \frac{1}{t+1}$. P_3 represents the segmentation inference with $f_{\text{ViT}}(\overline{\theta})$, which is used to supervise the feature-learning module following the consistency-aware concern. The supervision details of feature-learning ViT and CNN by guidance module are discussed in Section 3.4.

3.4 Objective

The training objective is to minimize the sum of the supervision loss $\mathcal{L}_{\text{sup}}$ and the semi-supervision $\mathcal{L}_{\text{semi}}$ among the three networks $f_{\text{CNN}}(\theta_1)$,$f_{\text{ViT}}(\theta_2)$, and $f_{\text{ViT}}(\overline{\theta})$, so the overall loss of S4CVnet being optimized during training is detailed in Equation 8:

$$\mathcal{L} = \mathcal{L}_{\text{sup1}} + \mathcal{L}_{\text{sup2}} + \lambda_1(\mathcal{L}_{\text{semi1}} + \mathcal{L}_{\text{semi2}}) + \lambda_2(\mathcal{L}_{\text{semi3}} + \mathcal{L}_{\text{semi4}}) \tag{8}$$

where λ_1, λ_2 are the weight factor of cross-supervision dual-view loss and consistency-aware loss, and it is updated every 150 iterations [26]. It is a trade-off weight that keeps increasing during the training process to make S4CVnet focus on labelled data when initialize, and then move focus to unlabeled data with our proposed semi-supervision approach. This is made under the assumption of the S4CVnet can gradually infer much reliable pseudo label confidently. The weight factor is briefly indicated in Equation 9.

$$\lambda = e^{-5\times(1-t_{\text{iteration}}/t_{\text{maxiteration}})^2} \tag{9}$$

where t indicates the current iteration number in a complete training process. Each of $\mathcal{L}_{\text{sup}}$ and $\mathcal{L}_{\text{semi}}$ are discussed as follows:

The semi-supervision loss among each network $\mathcal{L}_{\text{semi}}$ are calculated based on Cross-Entropy CE as shown in Equation 10:

$$\mathcal{L}_{\text{semi}} = \text{CE}\big(\text{argmax}(f_1(\boldsymbol{X};\theta), f_2(\boldsymbol{X};\theta))\big) \tag{10}$$

here we simply four $\mathcal{L}_{\text{semi}}$ losses with a pair of $\big(f_1(\boldsymbol{X};\theta), f_2(\boldsymbol{X};\theta)\big)$, where the pair can be $\big(f_{\text{CNN}}(\boldsymbol{X};\theta_1), f_{\text{ViT}}(\boldsymbol{X};\theta_2)\big)$, $\big(f_{\text{ViT}}(\boldsymbol{X};\theta_2), f_{\text{CNN}}\boldsymbol{X};\theta_1)\big)$, $\big(f_{\text{ViT}}(\boldsymbol{X};\overline{\theta}), f_{\text{ViT}}(\boldsymbol{X};\theta_2)\big)$, and $\big(f_{\text{ViT}}(\boldsymbol{X};\overline{\theta}), f_{\text{CNN}}(\boldsymbol{X};\theta_1)\big)$.

The supervision loss $\mathcal{L}_{\text{sup}}$ for each network is calculated based on both CE and the Dice Coefficient Dice as shown in Equation 11:

$$\mathcal{L}_{\text{sup}} = \frac{1}{2} \times \big(\text{CE}(Y_{\text{gt}}, f(\boldsymbol{X};\theta)) + \text{Dice}(Y_{\text{gt}}, f(\boldsymbol{X};\theta))\big) \tag{11}$$

Here we simply two $\mathcal{L}_{\text{sup}}$ supervision losses with a network $f(\boldsymbol{X};\theta)$, which can be considered as $f_{\text{CNN}}(\theta_1)$, and $f_{\text{ViT}}(\theta_2)$, because each network trained with labeled data Y_{gt} set is directly with the same way. The S4CVnet and all other baseline methods reported in Section 4 are with the same loss design including CE, and Dice for $\mathcal{L}_{\text{sup}}$ and $\mathcal{L}_{\text{semi}}$ in order to conduct a fair comparison.

4 Experiments and Results

Data set Our experiments validate the S4CVnet and all other baseline methods on the MRI ventricle segmentation data set from the automated cardiac diagnosis MICCAI Challenge 2017 [2]. The data is from 100 patients (nearly 6 000 images) 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. 20% of images are selected as the testing set, and the rest of the data set is for training(including validation).

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 four Nvidia GeForce RTX 3090 GPU, and Intel(R) Intel Core i9-10900K. The runtimes averaged around 5 hours, including the data transfer, training, inference and evaluation. The data set is processed for 2D image segmentation purposes. S4CVnet is trained for 30,000 iterations, the batch size is set to 24, the optimizer is SGD, and the learning rate is initially set to 0.01, momentum is 0.9, and weight decay is 0.0001. The network weight is saved and evaluated on the validation set every 200 iterations, and the network of guidance module with the best validation performance is used for final testing. The setting is also applied to other baseline methods directly without any modification.

Backbone The S4CVnet consists of two types of networks as shown in Figure 1. One is CNN-based segmentation network with skip connection, UNet[34], and the other one is ViT-based segmentation network with shift window[27] and skip connection, Swin-UNet[4]. For a fair comparison, two networks are both with U-shaped architecture with purely CNN- or ViT-based blocks as encoders and decoders. The tiny version of ViT is selected in this study to make the computational cost and training efficiency similar to CNN.

Baseline Methods All methods including S4CVnet and other baseline methods are trained with the same hyper-parameter setting, and the same distribution of features. The randomly selection of test set, labelled train set and unlabeled train set are only conducted once and then tested with all baseline methods together as well as S4CVnet. The baseline methods reported includes: MT [39], DAN [52], ICT [41], ADVENT [42], UAMT [51], DCN [32], CTCT [30] with CNN as the backbone segmentation network.

Evaluation Measures The direct comparison experiments between S4CVnet and other baseline methods are conducted with a variety of evaluation metrics including similarity measures: Dice, IOU, Accuracy, Precision, Sensitivity, and Specificity, which are the higher the better. The difference measures are also investigated: Hausdorff Distance (HD), and Average Surface Distance (ASD), which are the lower the better. 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 S4CVnet against other baseline methods, and the topological exploration of all alternative frameworks. IOU as the most common metric is also selected to report the performance of all baseline methods and S4CVnet under the assumption of different ratios of labelled data/total data. IOU, Sensitivity, and Specificity are selected to report the ablation study of different networks with different combinations of our proposed contribution.

Qualitative Results Figure 3 illustrates eight randomly selected sample raw images with related predicted images against the published ground truth, where Yellow, Red, Green and Black represent as True Positive(TP), False Positive(FP), False Negative(FN) and True Negative(TN) inferences at pixel level, respectively. This illustrates how S4CVnet can give rise to fewer FP pixels and lower ASD compared to other methods.

Quantitative Results Table 1 reports the direct comparison of S4CVnet against other semi-supervised methods including similarity measures and difference measures when the ratio of assumed labelled data/total data is 10%. The best result of different measures on the table is in **Bold**. A line chart in logarithmic scale is briefly sketched in Figure 4 (a), where the X-axis is the ratio of labelled data/total data, and Y-axis is the IOU performance, illustrating the valuable performance of S4CVnet against other baseline methods, especially in a low ratio of labelled data/total data. Details of quantitative results of S4CVnet and baseline methods under different assumption of ratio of labelled data/total data is in Appendix. A histogram indicating the cumulative distribution of IOU performance of prediction image is briefly sketched in Figure 4 (b), where the X-axis is the IOU threshold and the Y-axis is the number of predicted images on the test set, demonstrating S4CVnet is more likely to predict images with high IOU against other methods.

Framework	mDice↑	mIOU↑	Acc↑	Pre↑	Sen↑	Spe↑	HD↓	ASD↓
MT[39]	0.8860	0.8034	0.9952	0.8898	0.8829	0.9720	9.3659	2.5960
DAN[52]	0.8773	0.7906	0.9947	0.8721	0.8832	0.9743	9.3203	3.0326
ICT[41]	0.8902	0.8096	0.9954	0.8916	0.8897	0.9745	11.6224	3.0885
ADVENT[42]	0.8728	0.7836	0.9947	0.8985	0.8517	0.9601	9.3203	3.5026
UAMT[51]	0.8683	0.7770	0.9946	0.8988	0.8416	0.9582	8.3944	2.2659
DCN[32]	0.8809	0.7953	0.9951	0.8915	0.8714	0.9690	8.9155	2.7179
tri[10]	0.8918	0.7906	0.9947	0.8721	0.8832	0.9743	**7.2026**	2.2816
CTCT[30]	0.8998	0.8245	0.9959	0.8920	0.9083	0.9825	9.6960	2.7293
S4CVnet	**0.9146**	**0.8478**	**0.9966**	**0.9036**	**0.9283**	**0.9881**	12.5359	**0.6934**

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

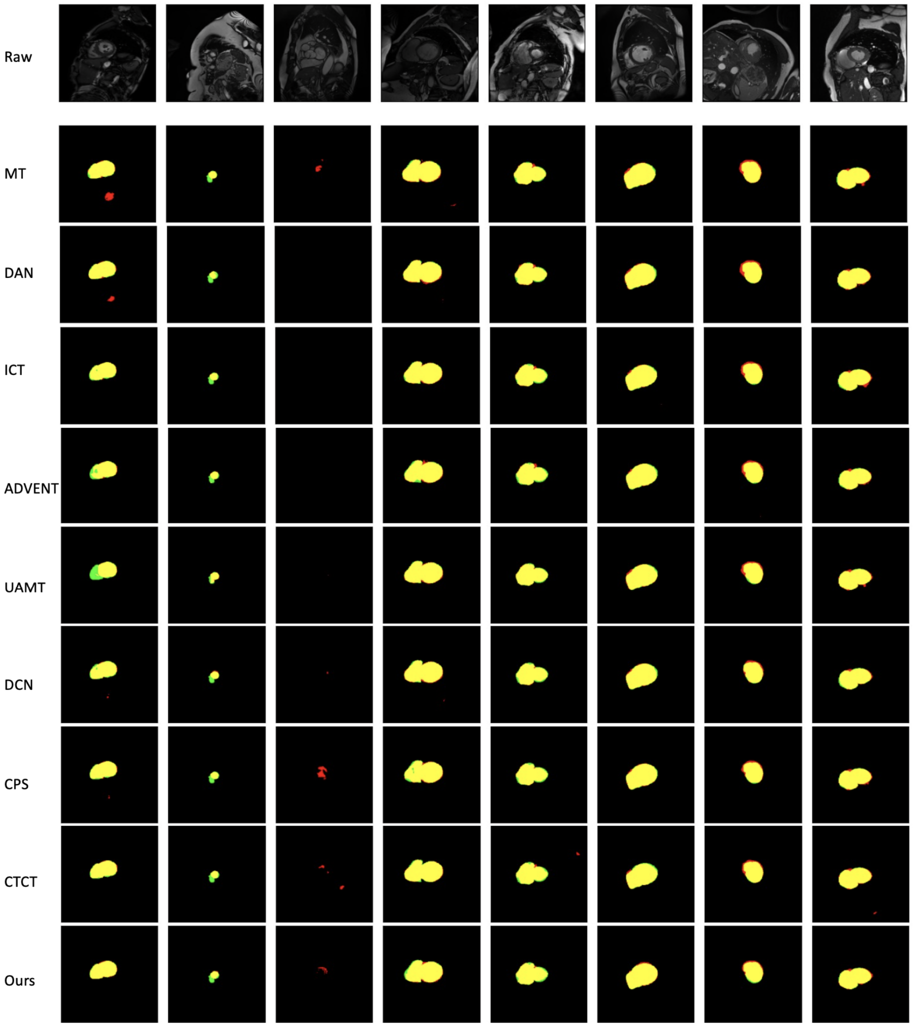


Fig. 3. Sample Qualitative Results on MRI Cardiac Test Set. Yellow, Red, Green, and Black Indicate True Positive, False Positive, False Negative, and True Positive of Each Pixel.

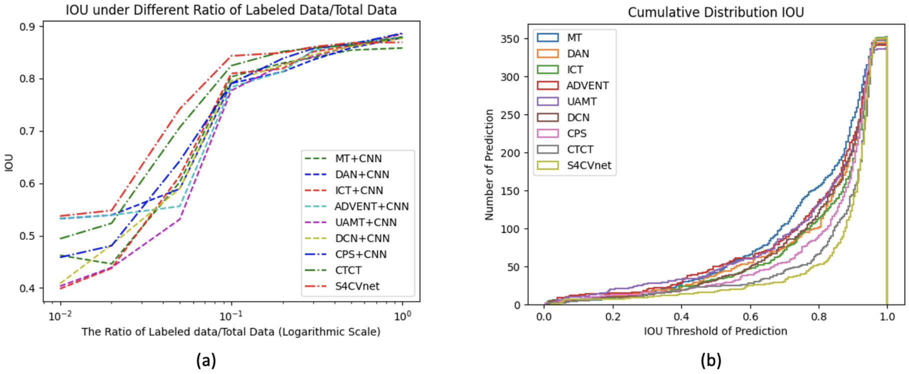


Fig. 4. The Performance of S4CVnet Against Other Baseline Methods. (a) The line chart of mIOU results on the test set with different assumptions of the ratio of label/total data for training. (b) The histogram chart indicates the cumulative distribution of IOU performance of the predicted image on the test set.

Ablation Study In order to analyze the effects of each of the proposed contributions and combinations including the setting of the network, the mechanism of the feature-learning module and guidance module, and the robustness of each network of S4CVnet, extensive ablation experiments have been conducted and reported in Table 2. ✗indicates either a network of feature-learning module or a guidance module is removed, and all alternative network settings(CNN or ViT) are explored. In different combinations of the proposed contribution, all the available networks are also tested separately, and the ablation study demonstrates that the proposed S4CVnet is with the most proper setting to fully utilize the power of CNN and ViT via student-teacher guidance scheme and dual-view co-training feature-learning approach in semi-supervised image semantic segmentation.

Supervision Mode Exploration Besides the ablation study to explore the different settings and combination of networks, feature-learning module, and guidance module, we fully explore the semi-supervised learning in medical image semantic segmentation through topological exploration of all alternative supervision modes of CNN and ViT. The full list of alternative frameworks is illustrated in Figure 5, where two supervision mode is briefly sketched in the legend of the figure. $\longrightarrow$ indicates the Student-Teacher style supervision mode, and $\dashrightarrow$ indicate cross pseudo-label-based supervision mode. Figure 5 (a) briefly illustrates all alternative frameworks with two axes, the Y-axis with different supervision modes from three networks to two networks, and the X-axis with the ratio of the number of CNN/ViT networks, and the proposed S4CVnet is in a red bounding box. All frameworks shown in the Figure 5 (a) have been tested and reported with heatmap format directly. Figure 5 (b) is an example heatmap to indicate the supervision mode and ratio of CNN/ViT information depending on the position of heatmap with a red bounding box to illustrate where is the S4CVnet as well. Figure 5 (c,d,e,f,g,h,i,j) represent the heatmap with mDice, mIOU, accuracy, precision, sensitivity, specificity, ASD, and HD validation performance, which demonstrate a whole picture of semi-supervised learning for medical semantic segmentation

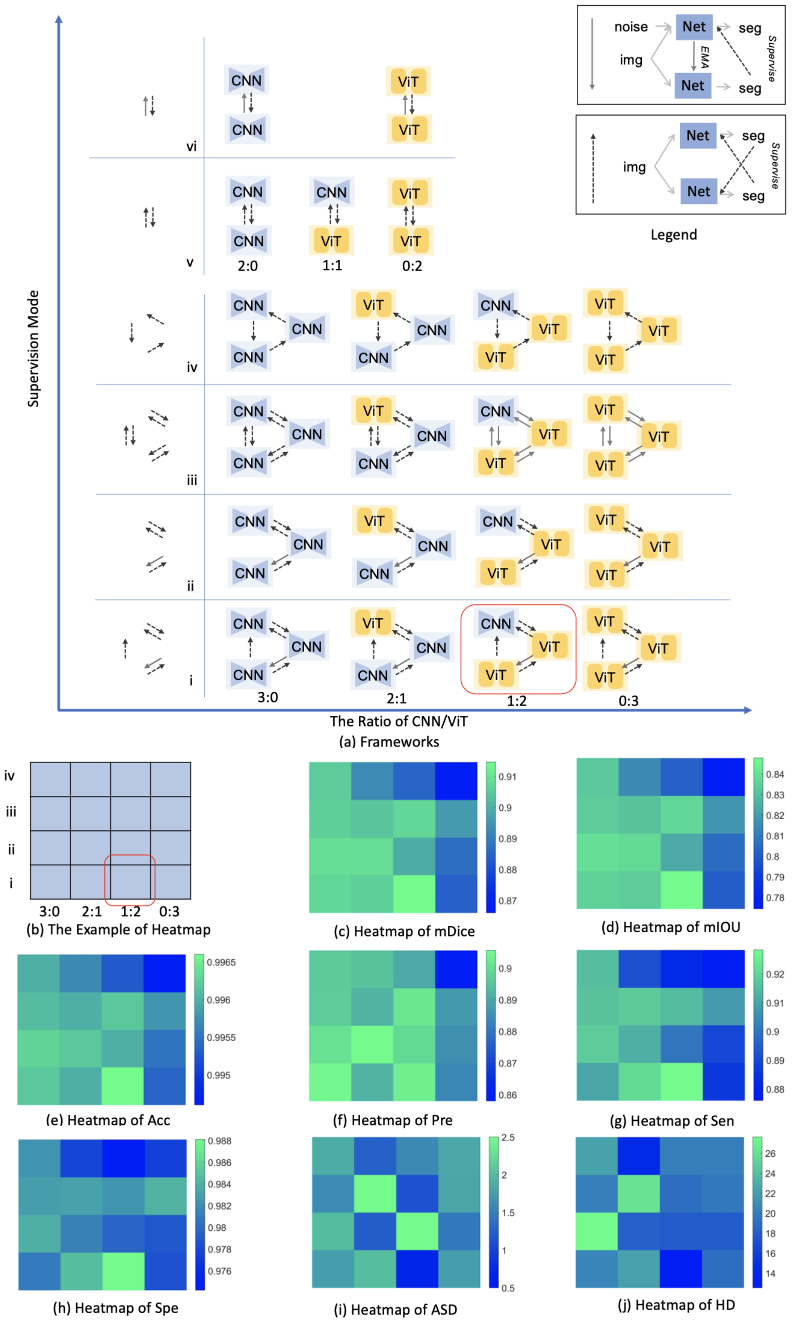


Fig. 5. The Topological Exploration of the Network(CNN&ViT), and Semi-Supervised Supervision Mode (Student-Teacher Style & Pseudo-Label).

Learning Module		Guidance Module	Test Network	IOU↑	Sen↑	Spe↑
Network A	Network B	Network C				
ViT	ViT	✗	A	0.8034	0.8829	0.9720
ViT	ViT	✗	B	0.8135	0.9036	0.9821
CNN	CNN	✗	A	0.7906	0.8832	0.9743
CNN	CNN	✗	B	0.8231	0.8967	0.9761
✗	CNN	CNN	B	0.7345	0.8094	0.9586
✗	CNN	CNN	C	0.7660	0.8481	0.9585
✗	ViT	ViT	B	0.8159	0.9032	0.9822
✗	ViT	ViT	C	0.7359	0.8415	0.9716
ViT	ViT	ViT	A	0.8096	0.8995	0.9817
ViT	ViT	ViT	B	0.8194	0.9078	0.9833
ViT	ViT	ViT	C	0.8183	0.9037	0.9822
CNN	CNN	CNN	A	0.8399	0.9225	0.9848
CNN	CNN	CNN	B	0.8432	0.9189	0.9848
CNN	CNN	CNN	C	0.8345	0.9168	0.9828
CNN	ViT	ViT	A	0.8341	0.9135	0.9825
CNN	ViT	ViT	B	0.8354	0.9177	0.9839
CNN	ViT	ViT	C	**0.8478**	**0.9283**	**0.9881**

Table 2. Ablation Studies on Contributions of Architecture and Modules

with CNN and ViT, and the denominating position of our proposed S4CVnet. The details of the quantitative results of topological exploration is in Appendix.

5 Conclusions

In this paper, we introduce an advanced semi-supervised learning framework in medical image semantic segmentation, S4CVnet, aiming to fully utilize the power of CNN and ViT simultaneously. S4CVnet consists of a feature-learning module and a guidance module. The feature-learning module, a dual-view feature learning approach, is proposed to enable two networks to complement each other via pseudo-label supervision. The guidance module is based on averaging network weights to supervise the learning modules under the consistency concern. Our proposed methods is evaluated with a variety of evaluation metrics and different assumption of the ratio of labelled data/total data against other semi-supervised learning baselines with the same hyperparameters settings and keeps the state-of-the-art position on a public benchmark data set. Besides a comprehensive ablation study, a topological exploration with CNN and ViT illustrates a whole picture of utilizing CNN and ViT in semi-supervised learning.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
2. Bernard, O., 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)
3. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the eleventh annual conference on Computational learning theory. pp. 92–100 (1998)
4. Cao, H., et al.: Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537 (2021)
5. Chang, Y.T., et al.: Weakly-supervised semantic segmentation via sub-category exploration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8991–9000 (2020)
6. Chen, J., etc: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
7. Chen, L.C., etc: Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In: European Conference on Computer Vision. pp. 695–714. Springer (2020)
8. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: 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)
9. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision. pp. 801–818 (2018)
10. Chen, X., etc: Semi-supervised semantic segmentation with cross pseudo supervision. In: CVPR (2021)
11. Deng, J., etc: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
12. Dong-DongChen, W., WeiGao, Z.H.: Tri-net for semi-supervised deep learning. In: Proceedings of twenty-seventh international joint conference on artificial intelligence. pp. 2014–2020 (2018)
13. Dosovitskiy, A., etc: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
14. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
15. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. Advances in neural information processing systems **31** (2018)
16. He, K., etc: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016)
17. Huang, B., et al: Simultaneous depth estimation and surgical tool segmentation in laparoscopic images. IEEE Transactions on Medical Robotics and Bionics **4**(2), 335–338 (2022)
18. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)

19. Hung, W.C., et al.: Adversarial learning for semi-supervised semantic segmentation. In: 29th British Machine Vision Conference, BMVC 2018 (2018)
20. Ibrahim, M.S., et al.: Semi-supervised semantic image segmentation with self-correcting networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12715–12725 (2020)
21. Ibtehaz, N., Rahman, M.S.: Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. Neural Networks **121**, 74–87 (2020)
22. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456. PMLR (2015)
23. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al.: nnu-net: Self-adapting framework for u-net-based medical image segmentation. In: Bildverarbeitung für die Medizin 2019, pp. 22–22. Springer (2019)
24. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Advances in neural information processing systems **28** (2015)
25. Ji, W., etc: Learning calibrated medical image segmentation via multi-rater agreement modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12341–12351 (2021)
26. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 (2016)
27. Liu, Z., et al: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
28. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
29. Luo, X., etc: Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2021)
30. Luo, X., etc: Semi-supervised medical image segmentation via cross teaching between cnn and transformer. arXiv preprint arXiv:2112.04894 (2021)
31. Mendel, R., Souza, L.A.d., Rauber, D., Papa, J.P., Palm, C.: Semi-supervised segmentation based on error-correcting supervision. In: European Conference on Computer Vision. pp. 141–157. Springer (2020)
32. Qiao, S., Shen, W., Zhang, Z., Wang, B., Yuille, A.: Deep co-training for semi-supervised image recognition. In: Proceedings of the european conference on computer vision. pp. 135–152 (2018)
33. Reiß, S., Seibold, C., Freytag, A., Rodner, E., Stiefelhagen, R.: Every annotation counts: Multi-label deep supervision for medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9532–9542 (2021)
34. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer (2015)
35. S Woo, e.a.: CBAM: Convolutional block attention module. In: Proc. pp. 3–19 (2018)
36. Song, C., et al.: Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3136–3145 (2019)

37. Souly, N., Spampinato, C., Shah, M.: Semi supervised semantic segmentation using generative adversarial network. In: Proceedings of the IEEE international conference on computer vision. pp. 5688–5696 (2017)
38. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7262–7272 (2021)
39. Tarvainen, A., etc: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in neural information processing systems (2017)
40. Vaswani, A., etc: Attention is all you need. In: Advances in neural information processing systems (2017)
41. Verma, V., etc: Interpolation consistency training for semi-supervised learning. In: International Joint Conference on Artificial Intelligence (2019)
42. Vu, T.H., etc: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2517–2526 (2019)
43. Wang, Z.: Deep learning in medical ultrasound image segmentation: A review. arXiv preprint arXiv:2002.07703 (2020)
44. Wang, Z., 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)
45. Wang, Z., etc: Triple-view feature learning for medical image segmentation. In: MICCAI Workshop on Resource-Efficient Medical Image Analysis (2022)
46. Wang, Z., Voiculescu, I.: Quadruple augmented pyramid network for multi-class covid-19 segmentation via ct. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC) (2021)
47. Wang, Z., et al.: Computationally-efficient vision transformer for medical image semantic segmentation via dual pseudo-label supervision. In: IEEE International Conference on Image Processing (ICIP) (2022)
48. Wang, Z., et al.: An uncertainty-aware transformer for mri cardiac semantic segmentation via mean teachers. Annual Conference on Medical Image Understanding and Analysis (2022)
49. Xia, Y., etc: 3d semi-supervised learning with uncertainty-aware multi-view co-training. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3646–3655 (2020)
50. You, X., et al.: Segmentation of retinal blood vessels using the radial projection and semi-supervised approach. Pattern recognition **44**(10-11), 2314–2324 (2011)
51. Yu, L., etc: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer (2019)
52. Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z.: Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: International conference on medical image computing and computer-assisted intervention. pp. 408–416. Springer (2017)
53. Zhou, B., et al.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)
54. Zoph, B., etc: Rethinking pre-training and self-training. Advances in neural information processing systems **33**, 3833–3845 (2020)

6 Appendix / Supplementary Material

6.1 Details of Experiments and Code

We used UNet and Swin-UNet as the segmentation backbone of CNN and ViT, respectively. All the baseline methods, proposed S4CVnet, and topological exploration of semi-supervised learning with CNN and ViT are developed with the same hyperparameter setting including optimizer, learning rate, batch size, and loss function. The feature distribution of labelled data set, unlabeled data set, validation data set, and the test data set is same for all methods in the experiment section.
The ViT backbone network is available at [2], all baseline methods is available at [3] without any modification. To reproduce these results, S4CVnet code is available at [4].

6.2 Details of Topological Exploration Results

In Section 4, a topological exploration with supervision modes and networks for all alternative semi-supervision frameworks in image semantic segmentation has been tested, and reported with heatmaps to give a whole picture of proposed methods. To clearly illustrate our topological exploration, all frameworks are simply sketched in Figure 6, and the full quantitative results of all frameworks are reported in Table 3 accordingly.

6.3 Details of Qualitative Results Under Assumption of Different Rate of Labeled/Total Data

Table 4 reports mIOU of S4CVnet and other baseline methods on test set under the different assumption of ratio of labeled data/total data for training, i.e. 1%, 2%, 5%, 10%, 20%, 30%, 50%, and 100%, respectively .

[2] https://github.com/HuCaoFighting/Swin-Unet

[3] https://github.com/HiLab-git/SSL4MIS

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

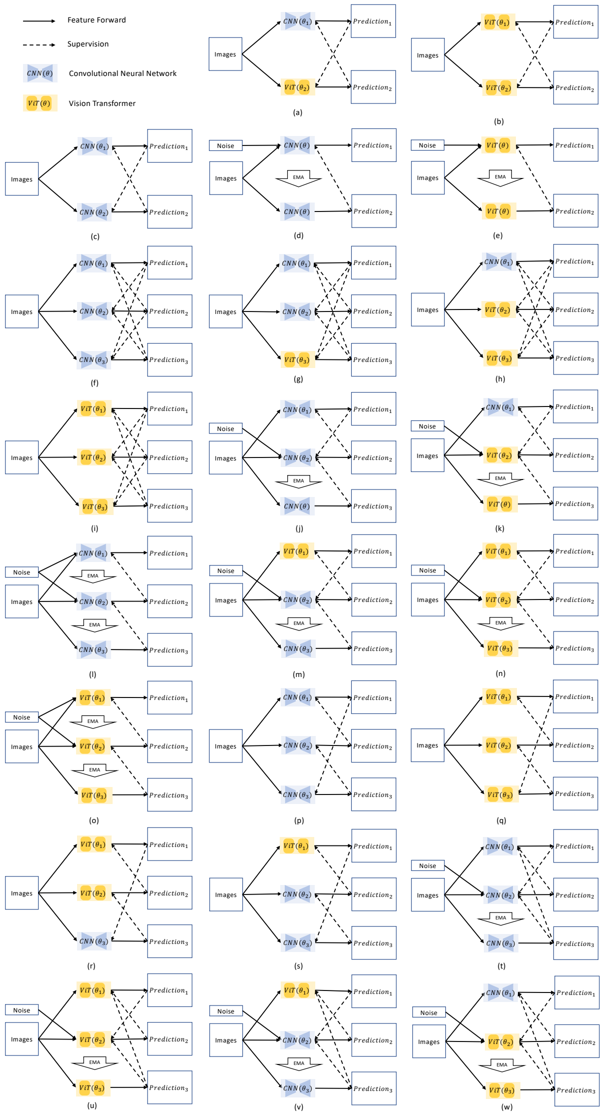


Fig. 6. The Full List of All Alternative Supervision Modes with CNN & ViT Frameworks with Our Proposed Techniques

Supervision Mode	mDice↑	mIOU↑	Acc↑	Pre↑	Sen↑	Spe↑	HD↓	ASD↓
A	0.8998	0.8245	0.9959	0.8920	0.9083	0.9825	9.6960	2.7293
B	0.8927	0.8135	0.9956	0.8832	0.9036	0.9821	17.7406	1.6316
C	0.8918	0.7906	0.9947	0.8721	0.8832	0.9743	**7.2026**	2.2816
D	0.8860	0.8034	0.9952	0.8898	0.8829	0.9720	9.3659	2.5960
E	0.8384	0.7359	0.9938	0.8361	0.8415	0.9716	23.7689	2.2801
F	0.9061	0.8341	0.9961	0.8970	0.9165	0.9829	20.1008	1.6110
G	0.9042	0.8311	0.9960	0.8918	0.9182	0.9831	26.2525	3.1882
H	0.9077	0.8372	0.9962	0.9022	0.9149	0.9825	19.1385	1.1296
I	0.8958	0.8184	0.9958	0.8866	0.9084	0.9841	19.7125	1.8150
J	0.9092	0.8391	0.9963	0.9006	0.9186	0.9838	27.6241	1.9961
K	0.8995	0.8243	0.9960	0.8993	0.9004	0.9797	17.9150	4.4550
M	0.9084	0.8380	0.9962	**0.9056**	0.9125	0.9814	18.2136	1.2208
N	0.8872	0.8054	0.9955	0.8856	0.8898	0.9788	17.9167	1.4451
P	0.9054	0.8330	0.9960	0.8965	0.9153	0.9823	22.2934	1.8733
Q	0.8660	0.7744	0.9946	0.8580	0.8760	0.9774	20.0766	1.8373
R	0.8854	0.8030	0.9953	0.8900	0.8819	0.9742	20.1741	1.5861
S	0.8930	0.8141	0.9957	0.8951	0.8919	0.9778	14.2738	1.2712
T	0.9074	0.8359	0.9962	0.9051	0.9104	0.9809	20.4208	1.7543
U	0.8843	0.8010	0.9954	0.8834	0.8860	0.9786	19.1564	1.7462
V	0.9060	0.8339	0.9960	0.8921	0.9212	0.9847	22.1800	2.0180
W(Ours)	**0.9146**	**0.8478**	**0.9966**	0.9036	**0.9283**	**0.9881**	12.5359	**0.6934**

Table 3. Direct Comparison of All Alternative Semi-supervised Frameworks with Our Proposed Techniques

Labeled/Total	1%	2%	5%	10%	20%	30%	50%	100%
MT[39]	0.4623	0.4460	0.6021	0.8034	0.8294	0.8397	0.8542	0.8583
DAN[52]	0.5323	0.5391	0.5892	0.7906	0.8130	0.8356	0.8585	0.8780
ICT[41]	0.3985	0.4376	0.6140	0.8096	0.8191	0.8512	0.8624	0.8853
ADVENT[42]	0.5329	0.5391	0.5559	0.7836	0.8133	0.8537	0.8677	0.8797
UAMT[51]	0.4034	0.4390	0.5310	0.7770	0.8269	0.8416	0.8619	0.8778
DCN[32]	0.4083	0.4824	0.5896	0.7953	0.8252	0.8455	0.8610	0.8769
CPS[10]	0.4583	0.4806	0.6426	0.7906	0.8383	0.8572	0.8654	**0.8865**
CTCT[30]	0.4939	0.5235	0.7066	0.8245	**0.8515**	0.8584	0.8638	0.8791
S4CVnet	**0.5374**	**0.5479**	**0.7418**	**0.8432**	0.8491	**0.8604**	**0.8679**	0.8691

Table 4. The Mean IOU Results on Test Set Under Different Assumptions of Ratio of Label/Total Data for Training