# Exigent Examiner and Mean Teacher: An Advanced 3D CNN-based Semi-Supervised Brain Tumor Segmentation Framework

Ziyang Wang and Irina Voiculescu

Department of Computer Science, University of Oxford, UK
{ziyang.wang, irina}@cs.ox.ac.uk

**Abstract.** With the rise of deep learning applications to medical imaging, there has been a growing appetite for large and well-annotated datasets, yet annotation is time-consuming and hard to come by. In this work, we train a 3D semantic segmentation model in an advanced semi-supervised learning fashion. The proposed SSL framework consists of three models: a *Student* model that learns from annotated data and a large amount of raw data, a *Teacher* model with the same architecture as the student, updated by self-ensembling and which supervises the student through pseudo-labels, and an *Examiner* model that assesses the quality of the student's inferences. All three models are built with 3D convolutional operations. The overall framework is a collaboration of consistency training Student $\leftrightarrow$ Teacher and adversarial training Examiner $\leftrightarrow$ Student. The proposed method is validated with various evaluation metrics on a public benchmarking 3D MRI brain tumor segmentation dataset. The experimental results of the proposed method outperform other state-of-the-art semi-supervised methods. The source code, 7 baseline methods, and dataset are available at https://github.com/ziyangwang007/CV-SSL-MIS.

**Keywords:** Semi-Supervised Learning, Image Semantic Segmentation, Mean Teacher, Adversarial Training, Brain Tumour Segmentation

## 1 Introduction

When applying deep-learning-based techniques to medical image segmentation, there has been a growing need for large and well-annotated datasets which, in turn, has come with a high cost in time and labor [1, 13, 17, 30, 26]. Semi-Supervised Learning (SSL) allows a model to be trained with a small amount of labeled data and a large amount of unlabeled data, has been widely explored in medical image analysis [18, 22, 33, 9, 29].

Consistency training is the most popular study in semantic segmentation with SSL. The consistency regularization is under the assumption when the perturbation is applied to the unlabeled data, the predictions should not change significantly, i.e. training a model with a consistent output. The perturbation

is explored normally with feature perturbation [16, 2, 18, 22], and network perturbation [11, 27]. A series of feature perturbations come with data augmentation such as CutMix[9] is developed to sufficiently varied perturbation, MixMatch[6] introduces low-entropy labels for data-augmented unlabeled examples and then mixes labeled and unlabeled data via MixUp[23], and FixMatch[12] utilises pseudo labels on weak augmentation to supervise the labeling on strong augmentation. On the other side, network perturbation studies normally come up with various network architectures such as dual-students[11] introducing stabilization constraint to an additional student for Student-Teacher style SSL [22], TriNet[7] explores three different decoders with a shared encoder and diversing data for classification, and triple-view learning[27] further explores multi-view learning for image semantic segmentation.

A self-ensembling SSL method named Mean Teacher [22], which is an extension of temporal ensembling [18], has been widely adopted in semi-supervised learning (SSL) for medical image segmentation [5, 28, 25, 19]. This method typically consists of a student model and a teacher model, with the same architecture. The student model learns from annotated data with feature perturbation, and the teacher model is updated by the average model weight from the student model. The teacher model is usually more robust than the student model and can supervise the student with pseudo-labels under a consistency-aware concern.

Alongside consistency training, another common SSL technique is adversarial training [15, 21, 10]. It normally involves developing an additional discriminator model to extract statistical features which aim to distinguish the quality of inferences of the model. Its standalone performance is still not serviceable in the clinical domain, especially not when the amount of labeled data is limited[8, 14].

Following the above concern, we explore the consistency-training-based SSL with further adversarial training scheme via extending the Student-Teacher style prototype with an Examiner paradigm, creating an Examiner↔Student↔Teacher SSL framework. To our best of knowledge, this is the first work that explore consistency training and adversarial training for medical image segmentation simultaneously with 3D convolutional operations. The contribution can be considered threefold: (i) The proposed framework consists of three models i.e. Examiner, Student, and Teacher, which are all based on 3D convolutions that can make the most of a training set which includes some annotated images as well as some unannotated raw data. (ii) Adversarial training and consistency regularization are proposed via Examiner ↔ Student, and Teacher ↔ Student respectively during the training process. (iii) Our framework has been validated on the public benchmark MRI brain tumor segmentation dataset[13] using comprehensive evaluation metrics. The results show that it outperforms seven other semi-supervised methods under the same hyper-parameter setting, segmentation backbone, and feature information distribution.

## 2 Approach

In this paper, $\mathbf{L}$, $\mathbf{U}$ and $\mathbf{T}$ denote a labeled training set, an unlabeled training set, and a test set. A batch of labeled training set is denoted as $(\boldsymbol{X}_{\mathrm{l}}, \boldsymbol{Y}_{\mathrm{gt}}) \in \mathbf{L}$, a batch of testing set as $(\boldsymbol{X}_{\mathrm{t}}, \boldsymbol{Y}_{\mathrm{gt}}) \in \mathbf{T}$, and unlabeled training set as $(\boldsymbol{X}_{\mathrm{u}}) \in \mathbf{U}$, where $\boldsymbol{X}_{\mathrm{l}}, \boldsymbol{X}_{\mathrm{t}}, \boldsymbol{X}_{\mathrm{u}} \in \mathbb{R}^{h \times w}$, and $\boldsymbol{Y}_{\mathrm{gt}} \in [0,1]^{h \times w}$ represent 2D grey-scale images, and their corresponding ground-truth annotations, respectively. A prediction $\boldsymbol{Y}_{\mathrm{p}} \in [0,1]^{h \times w}$ is generated by a segmentation model $f(\theta) : \boldsymbol{X} \mapsto \boldsymbol{Y}_{\mathrm{p}}$ using the parameters $\theta$ of the model $f$. This $\boldsymbol{Y}_{\mathrm{p}}$ can also be considered as a batch of unlabeled data with pseudo labels $(\boldsymbol{X}_{\mathrm{u}}) \in \mathbf{U}$, and the set of pairs $(\boldsymbol{X}_{\mathrm{u}}, \boldsymbol{Y}_{\mathrm{p}})$ can be used to retrain a model $f$. The Examiner, Student and Teacher are denoted by $f_E(\theta_{\mathrm{e}}), f_S(\theta_{\mathrm{s}}), f_T(\overline{\theta})$ respectively, where $\theta$ represents the parameters of each model, and $\overline{\theta}$ represents the average model weights constructed by the Teacher. The training of the Examiner $\leftrightarrow$ Student $\leftrightarrow$ Teacher composite minimizes the supervision loss $Loss_{\mathrm{sup}}$ and the semi-supervision loss $Loss_{\mathrm{semi}}$ of the Student and Examiner, respectively. $Loss_{\mathrm{sup}}$ and $Loss_{\mathrm{semi}}$ differ from each other through the use of genuine ground truth $\boldsymbol{Y}_{\mathrm{gt}}$ or generated pseudo labels $\boldsymbol{Y}_{\mathrm{p}}$. Our proposed SSL framework is briefly illustrated in Figure. 1, and the individual models $f_E(\theta_{\mathrm{e}})$, $f_S(\theta_{\mathrm{s}})$, and $f_T(\overline{\theta})$ are detailed in the following sections.

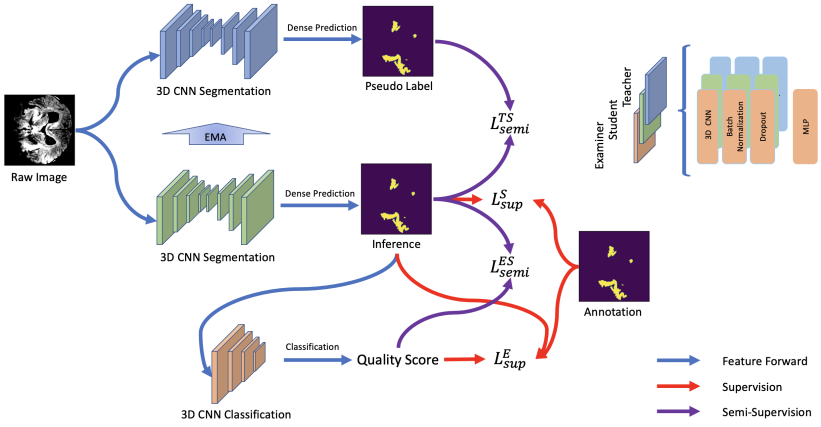


**Fig. 1.** The Framework of Examiner↔Student↔Teacher for Brain Tumor Segmentation.

### 2.1 Student Model

In order to exploit the 3D nature of the MRI scans, a 3DUNet [4] is used as the Student model $f_S(\theta)$. Each network block of this UNet is based on 3D convolutional operations, Batch Normalization and DropOut shown in Figure 1. Like

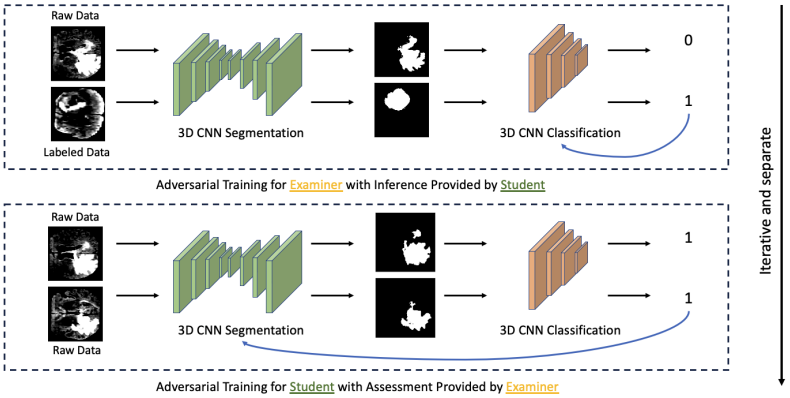


**Fig. 2.** The Illustration of Adversarial Training Between Examiner and Student.

in [22], $f_S(\theta)$ learns directly from data with annotations ($\boldsymbol{X}_{\mathrm{l}}$, $\boldsymbol{Y}_{\mathrm{gt}}$), and supervised by the Teacher via pseudo labels ($\boldsymbol{X}_{\mathrm{u}}$, $\boldsymbol{Y}_{\mathrm{p}}$). The crucial difference is that, at the same time, $f_S(\theta)$ is also validated against the Examiner via adversarial training. The inference of student model $f_S(\theta)$ is given in Eq. 1, where Gaussian noise is applied to all input data (both labeled and unlabeled) $\mathbf{X} = \boldsymbol{X}_{\mathrm{l}} \cup \boldsymbol{X}_{\mathrm{u}}$ during training.

$$\boldsymbol{Y}_{\mathrm{p}} = f_{\mathrm{s}}(\boldsymbol{X} + \mathit{Noise}; \theta_{\mathrm{s}}) \tag{1}$$

### 2.2 Teacher Model

The architecture of the Teacher $f_T(\overline{\theta})$ is similar to $f_S(\theta)$, except the Teacher does not learn from data directly. It is updated from the exponential moving average(EMA) (illustrated in Eq. 2) of $f_S(\theta)$, and it is more likely than the Student to predict the correct inference. Under the consistency-aware concern of inference from the same input data [22], the Teacher's predictions are considered as the pseudo labels to supervise the Student.

$$\overline{\theta} = \alpha\theta_{t-1} + (1-\alpha)\theta_t \tag{2}$$

where $\overline{\theta}$ is updated based on the Student parameter $\theta_t$ from the previous training step $t$; weight factor $\alpha = 1 - \frac{1}{t+1}$. The pseudo labels are generated by the Teacher without noise as:

$$\boldsymbol{Y}_{\mathrm{p}} = f_{\mathrm{T}}(X_{\mathrm{u}}; \overline{\theta}). \tag{3}$$

Thus the unlabeled training set $X_u$ can be utilized to train the Student with ($\mathbf{X}_{\mathrm{u}}$, $\boldsymbol{Y}_{\mathrm{p}}$).

### 2.3 Examiner Model

In order to capitalize on adversarial learning [33], a 3D-CNN-based discriminator is put in place to assess the quality of the Student's inference. This matches the

metaphor for an Examiner which checks the quality of the learning. The Examiner consists of four 3D CNN layers, a down-sampling operation, and multi-linear layers shown in Figure 1. Its architecture is following the classical VGGNet[20]. The Examiner and Student are trained against each other repeatedly for the duration of the training(seen in Fig.2). The Examiner classifies the quality of its input is indicated as Eq.4.

$$\boldsymbol{Y}_{\mathrm{e}} = f_{\mathrm{E}}(\boldsymbol{Y}_{\mathrm{p}}; \theta_{\mathrm{e}}) \tag{4}$$

where a segmentation mask predicted by Student $\boldsymbol{Y}_{\mathrm{p}}$ originating from ground-truth label $\boldsymbol{Y}_{\mathrm{gt}}$ is marked as a 1:*pass*, whereas a Student inference from pseudo label is marked as a 0:*fail* $Y_{\mathrm{e}} \in [pass, fail]$. As the adversarial training progresses, the student provides increasingly higher-quality inferences, the examiner provides an increasingly strict assessment.

### 2.4 Objective

The training objective of the proposed SSL approach is to minimize the sum of the supervision loss $\mathcal{L}_{\mathrm{sup}}$ and the semi-supervision loss $\mathcal{L}_{\mathrm{semi}}$ of two models $f_E(\theta), f_S(\theta)$ as shown in Eq. 5:

$$\mathcal{L} = \underbrace{\mathcal{L}^{S}_{\mathrm{sup}} + \mathcal{L}^{E}_{\mathrm{sup}}}_{sup} + \lambda \underbrace{(\mathcal{L}^{TS}_{\mathrm{semi}} + \mathcal{L}^{ES}_{\mathrm{semi}})}_{semi} \tag{5}$$

where the losses are split as $\mathcal{L}_{\mathrm{sup}}$ or $\mathcal{L}_{\mathrm{semi}}$ depending on whether the feeding data is from a labeled training set or not, and the loss is controlled by a ramp-up weight factor $\lambda$ [28]. $S, E, TS, ES$ indicate the loss of Student model, Examiner model, Student model with the help of Teacher, and the Student model with the assessment of Examiner, respectively. The overall segmentation loss is based on the Dice Coefficient, denoted Dice(.), and on Cross-Entropy CE(.). Loss $\mathcal{L}^{S}_{\mathrm{sup}}$ is used to train the Student with labeled data $(\boldsymbol{X}_{\mathrm{l}}, \boldsymbol{Y}_{\mathrm{gt}}) \in \mathbf{L}$, as per Eq. 6:

$$\mathcal{L}^{S}_{\mathrm{sup}} = \mathrm{CE}(Y_{\mathrm{gt}}, f_S(\boldsymbol{X}_l; \theta)) + \mathrm{Dice}(Y_{\mathrm{gt}}, f_S(\boldsymbol{X}_l; \theta)) \tag{6}$$

Loss $\mathcal{L}^{E}_{\mathrm{sup}}$ is used to train the Examiner with inferences from either the labeled or the unlabeled training set, illustrated in Eq. 7. The score $f_E(\theta) \in \{0, 1\}$, where 1 and 0 stand for *pass*, and *fail*.

$$\begin{aligned} \mathcal{L}^{E}_{\mathrm{sup}} = & \ \mathrm{CE}(f_E(f_S(\boldsymbol{X}_l; \theta_S); \theta_E), 1) \\ & + \mathrm{CE}(f_E(f_S(\boldsymbol{X}_u; \theta_S); \theta_E), 0) \end{aligned} \tag{7}$$

Loss $\mathcal{L}^{TS}_{\mathrm{semi}}$ is used to train the Student under the Teacher's supervision. A hybrid loss is utilized as shown in Eq. 8.

$$\begin{aligned} \mathcal{L}^{TS}_{\mathrm{semi}} = & \ \mathrm{CE}(f_S(X_u; \theta), f_T(\boldsymbol{X}_U; \theta)) \\ & + \mathrm{Dice}(f_S(X_u; \theta), f_T(\boldsymbol{X}_U; \theta)) \end{aligned} \tag{8}$$

Loss $\mathcal{L}^{ES}_{\mathrm{semi}}$ is used to train the Student under the 'exigent' assessment of Examiner via adversarial learning. The training objective of $\mathcal{L}^{ES}_{\mathrm{semi}}$ for $f_S(\theta)$ is

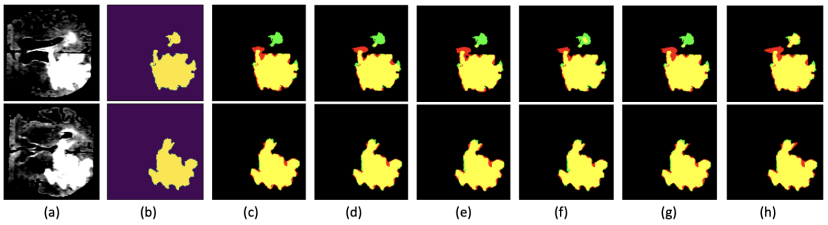


**Fig. 3.** Two sample images, corresponding annotations, and related inferences against annotations. (a)Raw image, (b)annotations, (c)ADVENT, (d)ICT, (e)UAMT, (f)CPS, (g)TVL, (h)ours.

for *all* the inferences to be *passed* by the Examiner. This makes the Examiner 'exigent' in that it repeats the examination process until all, or at least as many students as possible, have received a *pass*. In our experience, this can require as many as $O(10^4)$ iterations.

$$\mathcal{L}_{\text{semi}}^{ES} = -\text{CE}(f_E(f_s(\boldsymbol{X}_u; \theta_S); \theta_E), 0) \tag{9}$$

## 3 Experiments

**Implementation Details** The Examiner↔Student↔Teacher SSL method has been validated on the MRI Brain Tumor Segmentation(BraTS) from MICCAI Challenge 2019 [13]. It contains routine clinically-acquired 3T multimodal MRI scans, with accompanying ground-truth masks annotated by neuro-radiologists. We select flair images for whole tumor segmentation, with 80% selected for training (10% of which are labeled and the rest assumed unlabeled), and the remainder 20% for testing. The experiments are conducted with Pytorch on an Nvidia GeForce RTX 3090 GPU, and Intel(R) Intel Core i9-10900K. The 3D CNN segmentation is based on a modified 3DUNet [17, 4]. Runtimes average around 11.5 seconds per iteration. The dataset is processed for 3D semantic segmentation, with all images resized to 96×96×96. All baseline methods and our proposed method are trained with the same hyperparameter settings including training for 30,000 iterations then being tested with the model which had performed best on the validation set, batch size of 2, the optimizer for the Student is SGD, and the learning rate is initially set to 0.01 with 0.9 momentum and 0.0001 weight decay; the optimizer for the Examiner is Adam, and the learning rate is 0.0001. The Student with the best performance on validation set is for final testing.

**Compare with Baseline Methods** Our proposed method is compared with other SSL including Deep Adversarial Network (DAN) [33], Adversarial Entropy Minimization for Domain Adaptation (ADVENT) [24], Interpolation Consistency Training (ICT) [23], Mean Teachers (MT) [22], Uncertainty-Aware Mean Teachers (UAMT) [32], Cross Pseudo Supervision (CPS) [3], and Triple-View Learning (TVL) [27]. All the baseline SSL methods are with the *same* segmentation backbone i.e. 3DUNet for a fair comparison[4]. The comparisons are

**Table 1.** The Direct Comparison Between Each SSL Method on Brain Tumor MRI Testing Set When 10% of Training Set is Annotated.

| Model | Dice↑ | Acc↑ | Pre↑ | Sen↑ | Spe↑ | HD↓ | ASD↓ | SBD↑ |
|---|---|---|---|---|---|---|---|---|
| ADVENT[24] | 0.8458 | 0.9901 | 0.8935 | 0.8029 | 0.9967 | 12.8024 | 2.4089 | 0.5395 |
| ICT[23] | 0.8422 | 0.9900 | 0.8997 | 0.7917 | 0.9969 | 18.3787 | 2.5350 | 0.5399 |
| UAMT[32] | 0.8578 | 0.9908 | 0.8973 | 0.8217 | 0.9967 | 11.7392 | 2.4667 | 0.5547 |
| CPS[3] | 0.8580 | 0.9907 | 0.8882 | **0.8298** | 0.9963 | 12.9194 | **2.0330** | 0.5598 |
| TVL[27] | 0.8508 | 0.9903 | 0.8824 | 0.8214 | 0.9962 | 18.5677 | 2.2680 | 0.5955 |
| Ours | **0.8605** | **0.9911** | **0.9135** | 0.8134 | **0.9973** | **8.7455** | 2.1574 | **0.6013** |

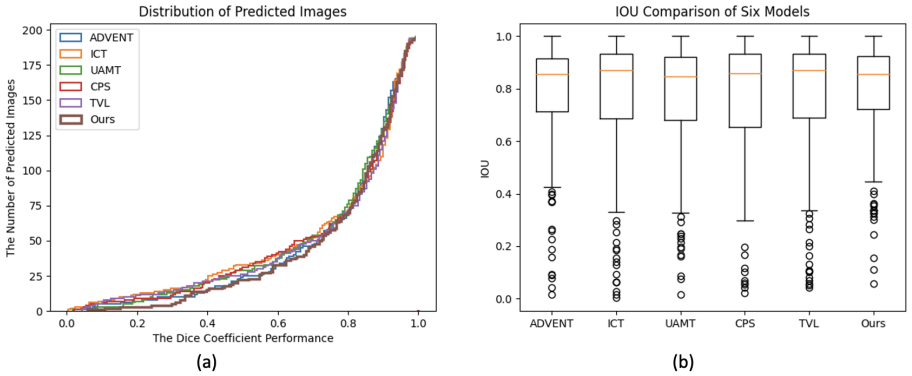


**Fig. 4.** The Dice and IoU Distribution of Each Inference by Six SSL Methods with Line and Box Chart.

conducted with a variety of evaluation metrics including Dice Coefficient (Dice), Accuracy (Acc), Precision (Pre), Sensitivity (Sen), Specificity (Spe), Hausdorff Distance (HD) 95% with millimeter, and Average Surface Distance (ASD). We further report a boundary-based metrics [31], using the Symmetric Boundary Dice (SBD). ↑ and ↓ denote the performance number(by similarity metrics or difference metrics) as the higher the better, or the lower the better, respectively.
**Results** The qualitative results are sketched in Figure. 3 illustrating two randomly selected raw images, annotations, and inference against the published annotations accordingly. Yellow, Red, Green and Black show true positive (TP), false positive (FP), false negative (FN) and true negative (TN) pixels when inference against published annotation. Some of the quantitative results are detailed in Table 1 with various metrics. The best performance are highlighted with **Bold**, and the second best of ours are highlighted with Underline. *It is worth noting that high Precision and low Sensitivity are not necessarily a good combination, nor are low Sensitivity and high Specificity.* Increasing Sensitivity (by 1.6% in Table 1) is desirable here as more of the tumour gets detected. Most of the results of our proposed methods are with **Bold** seen in Table 1. Except for reporting the average evaluation performance on the test set, each predicted image is also evaluated with Dice Coefficient seen in Figure. 4 (a) and (b). The

**Table 2.** The Direct Comparison Between Each SSL Method on Brain Tumor MRI Testing Set Under Different Data Situations.

| Labeled | 10% | | | 30% | | | 50% | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Dice↑ | HD↓ | ASD↓ | Dice↑ | HD↓ | ASD↓ | Dice↑ | HD↓ | ASD↓ |
| ADVENT[24] | 0.8458 | 12.8024 | 2.4089 | 0.8772 | 7.8178 | 2.1462 | 0.8769 | 8.1860 | 1.9881 |
| ICT[23] | 0.8422 | 18.3787 | 2.5350 | 0.8750 | 8.7348 | 1.9811 | 0.8741 | 7.3740 | 1.8393 |
| UAMT[32] | 0.8578 | 11.7392 | 2.4667 | 0.8793 | 7.6906 | 1.9216 | 0.8885 | 10.4160 | 1.7932 |
| CPS[3] | 0.8580 | 12.9194 | **2.0330** | 0.8770 | 7.2425 | **1.7739** | 0.8842 | 8.3445 | 1.8786 |
| TVL[27] | 0.8508 | 18.5677 | 2.2680 | 0.8798 | 11.1982 | 2.0076 | 0.8864 | 7.1704 | 1.8076 |
| Ours | **0.8605** | **8.7455** | 2.1574 | **0.8830** | **7.0310** | 1.8910 | **0.8920** | **6.9425** | **1.7628** |

line chart of Fig. 4 (a) illustrates the number of the predicted images on Y-Axis depending on the Dice threshold on X-Axis. The more curved the line, the more likely to predict segmentation with high Dice. The line of the best method(Ours) is highlighted with a **Thick** line. The box plot of Fig. 4 (b) directly illustrates the IOU distribution of predicted images of six different SSL methods with the same 3DUNet backbone. The qualitative and quantitative results both demonstrate the competitive performance of the proposed method against other SSL methods. The quantitative results shown in Table 1 are under the assumption that 10% of the training set is annotated. We further explored different data situations when 30% and 50% of the training set are with annotations shown in Table 2. **Ablation Study** To analyze the individual effects of each proposed contribution, as well as their combined effects, we conducted extensive ablation experiments, which are detailed in Table 3. All supervision schemes are with marks ✓on the mandatory Student model, because it is the only feature learning model from a limited annotation set directly. ✓with Teacher or Examiner indicates the only Teacher-Student consistency training or Examiner-Student adversarial training SSL scheme, which are both able to help the Student model learn from unannotated medical data. Further experiments of only fully supervised learning of the student model with 10% annotated data and 100% annotated data are also conducted as the lower-bound and upper-bound performance, respectively. In the lower-bound experiment, only the Student model is with a single ✓, which has no feature learning for raw data but with only 10% provided annotated data. In the upper-bound experiment, all three models are deployed with 100% of annotated data provided. Table 3 presents the promising improvement for Student with the help of Teacher and Examiner model.

## 4 Conclusion

Leveraging the power of adversarial training in the form of an Examiner model in conjunction with the Mean Teacher consistency regularization paradigm makes for a powerful combination which offers an improvement to semi-supervised learning for medical image segmentation. This has been validated on the widely used BraTS database but is easy to generalise to any semantic segmentation and, indeed, to other classes of downstream tasks.

**Table 3.** The Ablation Study with Two Separate Semi-Supervision Schemes and Fully Supervised Learning on Brain Tumor MRI Testing Set.

| Supervision | | | Performance | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| St | Te | Ex | Dice↑ | Acc↑ | Pre↑ | Sen↑ | Spe↑ | HD↓ | ASD↓ |
| ✓ (10% Full) | | | 0.8405 | 0.9898 | 0.8889 | 0.7970 | 0.9965 | 12.2822 | 2.2771 |
| ✓ | ✓ | | 0.8546 | 0.9906 | 0.8995 | 0.8139 | 0.9968 | 12.4331 | **2.1487** |
| ✓ | | ✓ | 0.8555 | 0.9906 | 0.8917 | **0.8222** | 0.9965 | 12.3674 | 2.5333 |
| ✓ | ✓ | ✓ | **0.8605** | **0.9911** | **0.9135** | 0.8134 | **0.9973** | **8.7455** | 2.1574 |
| 100% Full | | | 0.9073 | 0.9931 | 0.9130 | 0.9018 | 0.9968 | 7.4953 | 1.9241 |

## References


1. Bernard, O., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE TMI (2018)
2. Chen, L.C., et al.: Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In: ECCV (2020)
3. Chen, X., et al.: Semi-supervised semantic segmentation with cross pseudo supervision. In: CVPR (2021)
4. Cicek, et al.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: MICCAI (2016)
5. Cui, W., et al.: Semi-supervised brain lesion segmentation with an adapted mean teacher model. In: IPMI (2019)
6. David, B., et al.: Mixmatch: A holistic approach to semi-supervised learning. NeurIPS (2019)
7. Dong-DongChen, et al.: Tri-net for semi-supervised deep learning. In: IJCAI (2018)
8. Fang, K., Li, W.J.: DMNet: difference minimization network for semi-supervised segmentation in medical images. In: MICCAI (2020)
9. French, G., et al.: Semi-supervised semantic segmentation needs strong, varied perturbations. BMVC (2019)
10. Hung, W.C., et al.: Adversarial learning for semi-supervised semantic segmentation. In: BMVC (2018)
11. Ke, Z., et al.: Dual student: Breaking the limits of the teacher in semi-supervised learning. In: CVPR (2019)
12. Kihyuk, S., et al.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. NeurIPS (2020)
13. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BraTS). IEEE TMI (2014)
14. Mittal, S., Tatarchenko, M., Brox, T.: Semi-supervised semantic segmentation with high-and low-level consistency. IEEE TPAMI (2019)
15. Nasim, S., et al.: Semi supervised semantic segmentation using generative adversarial network. In: ICCV (2017)
16. Ouali, Y., et al.: Semi-supervised semantic segmentation with cross-consistency training. In: CVPR (2020)
17. Ronneberger, O., et al.: U-net: Convolutional networks for biomed image segmentation. In: MICCAI (2015)
18. Samuli, L., Aila, T.: Temporal ensembling for semi-supervised learning. ICLR (2016)

19. Shanis, Z., et al.: Intramodality domain adaptation using self ensembling and adversarial training. In: MICCAI. Springer (2019)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. ICLR (2015)
21. Takeru, M., et al.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE TPAMI (2018)
22. Tarvainen, A., et al.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NeurIPS (2017)
23. Verma, V., et al.: Interpolation consistency training for semi-supervised learning. In: IJCAI (2019)
24. Vu, T.H., et al.: ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: CVPR (2019)
25. Wang, K., et al.: Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. MedIA (2022)
26. Wang, Z., et al.: RAR-U-net: a residual encoder to attention decoder by residual connections framework for spine segmentation under noisy labels. In: ICIP. IEEE (2021)
27. Wang, Z., et al.: Triple-view feature learning for medical image segmentation. In: MICCAI-W (2022)
28. Wang, Z., et al.: Uncertainty-aware transformer for MRI cardiac segmentation via mean teachers. MIUA (2022)
29. Wang, Z., Dong, N., Voiculescu, I.: Computationally-efficient vision transformer for medical image semantic segmentation via dual pseudo-label supervision. In: ICIP. IEEE (2022)
30. Wang, Z., Voiculescu, I.: Dealing with unreliable annotations: A noise-robust network for semantic segmentation through a transformer-improved encoder and convolution decoder. Applied Sciences (2023)
31. Yeghiazaryan, V., et al.: Family of boundary overlap metrics for the evaluation of medical image segmentation. SPIE JMI (2018)
32. Yu, L., et al.: Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: MICCAI (2019)
33. Zhang, Y., et al.: Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: MICCAI (2017)

## 5 Changes after Reviews

We thank Reviewer 1, 2, 3 and meta-reviewer for your positive comments and constructive suggestions. We have made some changes for the camera-ready paper and the responds have been splited as writing-related or experiment-related changes and discussed as following:

**Writing-related changes:**

(1) Reviewer 1: The summary of the contributions of the examiner-student-teacher has been improved in the end of introduction section. They are now discussed with several bullet points. (2) Reviewer 1: We have added some sentences to detailed discuss the meaning of 10% full and 100% full experiments in the ablation study section. (3) Reviewer 2: We have redraw Figure 4(a). In this figure, we illustrate the number of prediction according to the dice performance. The dice performance of each prediction should be between 0 and 1. By doing so, when dice is 1 in x-axis, the number of prediction should be equal to the total number of images on test set. In other words, when dice is 0 in x-axis, there should be no number of prediction. (4) Reviewer 2: This is the first work that explore adversaria training with consistency regularization for medical image segmentation. Considering 'The improvement id DICE is minimal', it can be well addressed if we validate with more dataset, or other backbone network such as Vision Transformer. In this study, we only use very basic 3D UNet, or 3D VGG net as segmentation backbone network. (5)Reviewer 3: We have explained the 'Quality Score' in the end of Section 2.3 Examiner Model. (6)Reviewer 3: Thank you for your detailed proofreading. We have corrected for Table 1. (7)Reviewer 3: In the past, we make the length of Figure 4 with 0.8 line length. We have made Figure 4 bigger with full line length, so that it should be much helpful for reader notice some difference between each method.

**Experiment-related changes:**

We thank Reviewer 1 suggested we should add more baseline methods. Reviewer 2 suggested we should evaluate on more data set not just BRATS -19. Reviewer 3 suggested we should add more experiments under different data situations, such as what if 30% or 50% dataset are with labeled. But considering the page limit and the tight time for preparing camera-ready paper after decision notification, we achieved all these suggestions on GitHub.

This project has been made available on https://github.com/ziyangwang007/CV-SSL-MIS.

(1) Reviewer 1 - More baseline methods: We have added more baseline methods available, such as mean teacher, adversarial learning, and etc. (2) Reviewer 2 - More datasets: we have added one more dataset about MRI Cardiac Segmentation Dataset. Same with BRATS-19 from MICCAI Chanllenge, the new dataset is ACDC from MICCAI Chanllenge. (3) Reviewer 3 - More data conditions: we have added more hyparameters setting in GitHub that you can choose 1%, 3%, 5%, 10%, 20%, 30%, 50%, 100% labeled data for experiments.

All these experiment-related changes have been completed, and please kindly check on the GitHub project page.