# Efficient MRI image segmentation using semi-supervised learning

Shihan Cheng[1] and Chang-Yong Song[1]

Vanderbilt University, Nashville TN 37235, USA
shihan.cheng@vanderbilt.edu, chang-yong.song@vanderbilt.edu

**Abstract.** Medical image segmentation often faces challenges in acquiring labeled data due to the time and cost involved in expert annotation. This study addresses these challenges using semi-supervised learning, enhanced by SimpleITK-based preprocessing to improve data quality and focus on critical structures. By comparing U-Net, Mean Teacher, and UAMT models, the results demonstrate that UAMT achieves superior performance in complex structures through consistency regularization and uncertainty-based pseudo-label filtering. This research highlights the potential of semi-supervised learning for accurate and efficient medical image segmentation, offering promising applications in clinical settings. https://github.com/csh-apprentice/CS6357_final

**Keywords:** Data preprocessing · SimpleITK · Semi-supervised learning

## 1 Introduction

Labeling medical imaging data requires precise annotations by experts, which makes acquiring large-scale labeled datasets time-consuming and expensive. This challenge is a major obstacle in obtaining sufficient training data for medical image analysis. However, incorporating unlabeled data into the training process offers the potential for models to learn more comprehensive representations of anatomical structures and their variations.

In this study, we utilize semi-supervised learning, which integrates limited labeled data with large amounts of unlabeled data for training. Specifically, we employ SimpleITK for preprocessing to enhance data quality and enable the model to focus on critical structures. This preprocessing pipeline contributes to improving image quality by normalizing intensity and contrast, removing irrelevant regions, and reducing noise.

We aim to systematically compare semi-supervised learning models with SimpleITK-based preprocessing against existing semi-supervised approaches. Through this comparison, we analyze the impact of preprocessing on model performance and generalization capabilities. Ultimately, our goal is to achieve accurate and efficient segmentation of medical images, thereby increasing the applicability of these methods in real-world clinical settings.

## 2 Related Work

Medical image segmentation encompasses a variety of models addressing challenges such as data scarcity and domain shifts. These include fully supervised models, such as UNet [9] and nnUNet [5], which require extensive labeled datasets, and unsupervised methods like Deep Image Prior [11], which rely on generative or self-supervised techniques. Semi-supervised learning bridges these approaches by leveraging both labeled and unlabeled data, exemplified by models like Mean Teacher [10] and Uncertainty-Aware Mean Teacher (UAMT) [7], which achieve a balance between accuracy and data efficiency.

We will focus on semi-supervised segmentation techniques, as they provide an effective framework for addressing the challenges of limited annotations while achieving competitive segmentation performance. Semi-supervised segmentation methods demonstrate robust performance even with limited labeled data. Key approaches include consistency-based methods, such as Mean Teacher and Uncertainty Rectified Pyramid Consistency (URPC) [7]; adversarial-based approaches, like Deep Adversarial Networks [4]; and pseudo-labeling techniques, such as Cross Pseudo Supervision (CPS). Additionally, innovations like uncertainty-aware models, such as UAMT, and transformer-based methods, including Cross Teaching Between CNN and Transformer [6], highlight the adaptability of semi-supervised learning in medical imaging. These strategies effectively address challenges related to annotation scarcity and model generalization.

## 3 Models

### 3.1 UNet3D Baseline

The UNet3D model [3], implemented as a fully convolutional neural network, is specifically designed for volumetric segmentation of 3D medical images. The architecture adopts an encoder-decoder structure with skip connections, enabling the integration of spatial features from the encoder path with detailed localization information from the decoder path. The encoder progressively reduces spatial dimensions through a series of downsampling operations, extracting hierarchical features, while the decoder restores spatial resolution using upsampling layers. At the network's core, a bottleneck layer captures high-level semantic information, acting as a transition between the encoder and decoder.

Our implementation of UNet3D employs a consistent number of filters, specifically $[16, 32, 64, 128, 256]$, across its encoder-decoder structure. Each encoder level consists of 3D convolutional layers with a kernel size of $(3, 3, 3)$ and padding of $(1, 1, 1)$, followed by optional batch normalization and 3D max-pooling with a kernel size of $(2, 2, 2)$. The bottleneck layer applies similar 3D convolutions, complemented by dropout ($p = 0.3$) for regularization. In the decoder, upsampling layers restore spatial dimensions and concatenate their outputs with features from corresponding encoder layers via skip connections. The final output is produced through a $1 \times 1 \times 1$ convolutional layer, generating voxel-wise predictions for segmentation.

### 3.2 Mean Teacher Model

The Mean Teacher model [10] is a semi-supervised learning framework that leverages a teacher-student paradigm for training. In this setup, two models are maintained: the student model, which is updated using standard backpropagation, and the teacher model, which is updated as an exponential moving average (EMA) of the student's weights. This approach enables the teacher to provide stable targets for unlabeled data, thus improving generalization. The student model is trained on both labeled and unlabeled data, with the teacher's predictions acting as pseudo-labels for the unlabeled samples. In our implementation, both the student and teacher models adopt the UNet3D architecture with the same UNet3D structure introduced in the last section.

The total loss function for the Mean Teacher model combines supervised loss on labeled data and consistency loss on unlabeled data to leverage both types of data effectively. The supervised loss, computed over a batch of labeled samples of size $B$, combines the Cross-Entropy Loss and Dice Loss as:

$$\mathcal{L} = \cdot\frac{1}{B}\sum_{i=1}^{B}\left(\mathcal{L}_{\text{CE}}(p_i, y_i) + \mathcal{L}_{\text{Dice}}(p_i, y_i)\right) + w(t)\cdot\frac{1}{U}\sum_{j=1}^{U}\left\|p_j^{\text{student}} - p_j^{\text{teacher}}\right\|^2. \quad (1)$$

Here, $p_i$ and $y_i$ represent the predicted probabilities and ground truth labels for labeled samples, while $p_j^{\text{student}}$ and $p_j^{\text{teacher}}$ are the predictions of the student and teacher models, respectively, for the unlabeled samples in a batch of size $U$. The consistency loss ensures that the student's predictions align with the teacher's predictions, weighted by a time-dependent factor $w(t)$ that increases over training following a sigmoid ramp-up schedule. This combined loss allows the model to benefit from labeled data for supervised learning while leveraging unlabeled data through consistency regularization, thereby enhancing segmentation performance.

### 3.3 Uncertainty Aware Mean Teacher Model (UAMT)

The Uncertainty-Aware Mean Teacher (UAMT) model builds upon the Mean Teacher framework by incorporating uncertainty estimation to enhance the utilization of unlabeled data. Similar to the Mean Teacher model, UAMT maintains a student-teacher paradigm. However, UAMT extends the Mean Teacher approach by dynamically weighting the consistency loss for unlabeled samples based on uncertainty, effectively filtering out unreliable pseudo-labels.

In UAMT, uncertainty is quantified using Monte Carlo Dropout. For each unlabeled sample, the teacher model generates multiple stochastic predictions by introducing random noise to the inputs during forward passes. These predictions are averaged to compute a probabilistic consensus, and the uncertainty is estimated as the entropy of the mean prediction:

$$\text{Uncertainty} = -\sum_{c=1}^{C} p_c \log(p_c), \quad (2)$$

where $p_c$ represents the predicted probability for class $c$. A thresholding mechanism is applied, where only the pseudo-labels with uncertainty below a dynamic threshold are incorporated into the consistency loss. The total loss function for UAMT is given by:

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + w(t) \cdot \frac{\sum_{j=1}^{U} 1(\text{Uncertainty}_j < \text{Threshold}) \cdot \mathcal{L}_{\text{cons}}(p_j^{\text{student}}, p_j^{\text{teacher}})}{\sum_{j=1}^{U} 1(\text{Uncertainty}_j < \text{Threshold}) + \epsilon}, \tag{3}$$

where $\mathcal{L}_{\text{sup}}$ is the supervised loss on labeled data, $\mathcal{L}_{\text{cons}}$ is the consistency loss, $w(t)$ is a time-dependent consistency weight, and 1 is the indicator function selecting pseudo-labels with uncertainty below the threshold. By explicitly accounting for uncertainty, UAMT mitigates the influence of noisy or incorrect pseudo-labels that could arise in challenging segmentation tasks. This makes it particularly effective in handling the inherent noise in medical imaging data. We shall see both the qualitive and quantity comparisions in the **Results Section**.

## 4 Training

### 4.1 Training Data

We use the **BraTS 2019 dataset** [8] [1], [2], available on Kaggle[1], for our experiments. Specifically, we utilize the FLAIR modality and its corresponding segmentation labels. From the dataset, which comprises 250 image-label pairs, we randomly select 25 pairs as labeled data. While the UNet3D model is restricted to training on this labeled subset, the Mean Teacher and UAMT models leverage both the labeled 25 pairs and the remaining 225 unlabeled pairs during their training, allowing them to effectively utilize the unlabeled data for semi-supervised learning.

### 4.2 Prepossessing

For effective training, we also have to Prepossess our data. First, we extract a bounding box around the brain region using a binary mask. Next, we normalize the intensity values within the brain region by calculating the mean and standard deviation of non-zero voxels and scaling the intensities to have zero mean and unit variance. To handle outliers, we apply intensity clipping based on the cumulative distribution function (CDF) with a pre-defined percentile threshold. Additionally, ground truth segmentation masks are binarized to create a consistent format for training segmentation models. All of these parts are implemented using **SimpleITK**. After that, we are applying augmentation by randomly rotate and clip the images.

[1] https://www.kaggle.com/datasets/aryashah2k/brain-tumor-segmentation-brats-2019

### 4.3 Optimizer

In this project, we use the Stochastic Gradient Descent (SGD) optimizer with momentum for training all models. The learning rate is scheduled to decrease over iterations using a polynomial decay function: $\text{lr} = \text{base\_lr} \cdot (1 - \frac{\text{iter}}{\text{max\_iter}})^{0.9}$, where base_lr is the initial learning rate. For weight initialization, we employ the Kaiming initialization strategy.

### 4.4 Metrics

In our experiments, we evaluate the segmentation performance using four commonly used metrics: Dice Similarity Coefficient (Dice), Relative Absolute Volume Difference (RAVD), 95th Percentile Hausdorff Distance (HD95), and Average Surface Distance (ASD).

## 5 Results

We conducted both training and testing on an Intel-Xeon CPU and NVIDIA A4000 GPU experimental setup, using training data descibed in previous section.

### 5.1 Training loss

The training loss curves presented in Fig. 1 highlight the performance differences between **U-Net**, **Mean Teacher**, and **UAMT** models. All models show a rapid decrease in total loss during the initial iterations, with the loss stabilizing as training progresses. Both UAMT and Mean Teacher maintain consistently lower loss values compared to U-Net, demonstrating superior learning stability and efficiency.

In terms of cross-entropy loss, UAMT and Mean Teacher stabilize faster and achieve lower final loss values, further demonstrating their effectiveness in optimizing classification boundaries. Dice loss trends reveal that while all three models follow a similar pattern after the initial decline, UAMT consistently achieves the lowest loss, showcasing its robustness in segmenting complex regions. However, UAMT exhibits a temporary spike in loss during certain iterations, likely due to the impact of consistency regularization, which enforces consistency in predictions for unlabeled data. This phenomenon may arise as the consistency loss term affects model updates, reflecting a characteristic of semi-supervised learning where the model may experience temporary instability while reducing structural uncertainty in the data.

These observations suggest that UAMT and Mean Teacher outperform U-Net in generalization to training data under preprocessed semi-supervised learning conditions. In particular, consistency regularization plays a key role in enhancing the utilization of unlabeled data and improving UAMT's segmentation performance in complex regions.

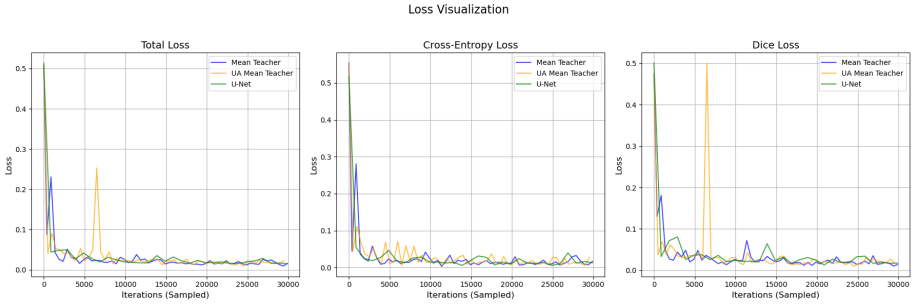



Fig. 1: Training loss curves for U-Net(**green**), Mean Teacher(**Blue**), and UAMT(**yellow**) models. The loss metrics include Total Loss, Cross-Entropy Loss, and Dice Loss, showing changes across sampled iterations.

### 5.2 Metrics analysis

The training loss curves presented in Fig. 1 highlight the performance differences between **U-Net**, **Mean Teacher**, and **UAMT** models. All models show a rapid decrease in total loss during the initial iterations, with the loss stabilizing as training progresses. Both UAMT and Mean Teacher maintain consistently lower loss values compared to U-Net, demonstrating superior learning stability and efficiency.

In terms of cross-entropy loss, UAMT and Mean Teacher stabilize faster and achieve lower final loss values, further demonstrating their effectiveness in optimizing classification boundaries. Dice loss trends reveal that while all three models follow a similar pattern after the initial decline, UAMT consistently achieves the lowest loss, showcasing its robustness in segmenting complex regions. However, UAMT exhibits a temporary spike in loss during certain iterations, likely due to the impact of consistency regularization, which enforces consistency in predictions for unlabeled data. This phenomenon may arise as the consistency loss term affects model updates, reflecting a characteristic of semi-supervised learning where the model may experience temporary instability while reducing structural uncertainty in the data.

These observations suggest that UAMT and Mean Teacher outperform U-Net in generalization to training data under preprocessed semi-supervised learning conditions. In particular, consistency regularization plays a key role in enhancing the utilization of unlabeled data and improving UAMT's segmentation performance in complex regions.

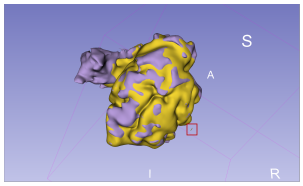
### 5.3 Visualization comparison

The 3D and 2D visualization comparisons revealed clear performance differences among U-Net, Mean Teacher, and UAMT. In the 3D visualization, U-Net exhibited a lack of structural accuracy, with yellow regions indicating its limitations in reproducing the boundaries and details of complex structures.

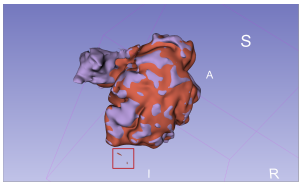
| **Metric** | U-Net | Mean teacher | UAMT |
|---|---|---|---|
| Mean Dice Similarity Coefficient (Dice) | 0.821 | 0.816 | 0.812 |
| Mean Relative Absolute Volume Difference (RAVD) | 0.231 | 0.230 | 0.253 |
| Mean 95th Percentile Hausdorff Distance (HD95) | 12.42 | 11.436 | 14.27 |
| Mean Average Surface Distance (ASD) | 3.31 | 3.002 | 3.256 |

Table 1: Summarize the performance metrics for U-Net, Mean Teacher, and UAMT models

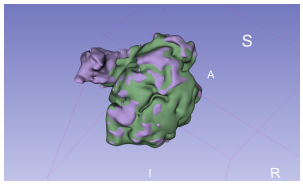
Mean Teacher showed improved alignment with the GT compared to U-Net, but some boundary inaccuracies, marked in red, were still present. In contrast, UAMT demonstrated the highest consistency and alignment with the GT, as evidenced by the green regions, accurately capturing boundaries and details. In the 2D axis-based comparisons, UAMT consistently outperformed other models in boundary and internal structure alignment across all axes, while Mean Teacher showed better similarity than U-Net along the Y and Z axes. U-Net, however, displayed boundary inaccuracies and distorted regions, resulting in relatively lower performance.



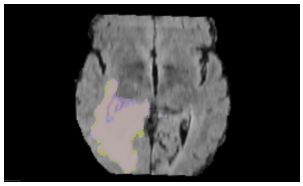

(a) Compare GT(purple) and U-Net(yellow).



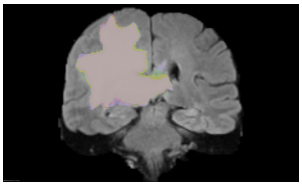

(b) Compare GT(purple) and Mean teacher(red).



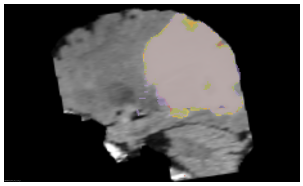

(c) Compare GT(purple) and UAMT(green).



(d) Compare all results based on the X-axis.



(e) Compare all results based on the Y-axis.



(f) Compare all results based on the Z-axis

Fig. 2: 3D and 2D Visualization Results. (a), (b), and (c) present the 3D visual comparison between the Ground Truth (GT, **purple**) and the results from U-Net (**yellow**), Mean Teacher (**red**), and UAMT (**green**), respectively. (d), (e), and (f) illustrate the 2D slice comparisons of all results across the X, Y, and Z axes.

### 5.4 Conclusion

This study proposed a method to enhance the performance and generalization capability of semi-supervised learning models by utilizing preprocessing steps based on SimpleITK and conducted a systematic comparison with existing semi-supervised learning approaches. By effectively integrating limited labeled data with a large amount of unlabeled data, the study demonstrated the potential of models to learn anatomical structures and variations comprehensively. The results confirmed that SimpleITK-based preprocessing contributed to improving data quality and enabling models to focus on critical structures, thereby positively impacting the accuracy and efficiency of medical image segmentation. These findings suggest that the proposed approach can further enhance the applicability of semi-supervised learning methods in real clinical environments.

## References

1. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data **4**(1), 1–13 (2017)
2. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018)
3. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. pp. 424–432. Springer (2016)
4. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
5. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
6. Luo, X., Hu, M., Song, T., Wang, G., Zhang, S.: Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In: International conference on medical imaging with deep learning. pp. 820–833. PMLR (2022)
7. Luo, X., Wang, G., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Metaxas, D.N., Zhang, S.: Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. Medical Image Analysis **80**, 102517 (2022)
8. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
9. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)

10. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
11. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Deep image prior. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9446–9454 (2018)