3D Semi-Supervised Segmentation on Left-Atrium MRIs with Cross Pseudo Supervision and Mean Teacher model with Uncertainty Map

GUPTA Pranav[1], GUO Jiarong[1]
Electronic and Computer Engineering Department,
The Hong Kong University of Science and Technology, Hong Kong[2]
Email: {pguptaaf@connect.ust.hk, jguoaz@connect.ust.hk}

I. Introduction:

3D Semi-supervised Learning for segmentation tasks has been studied well since the flourishing of deep learning technology. 3D Segmentation is useful in various applications such as autonomous vehicles, 3D modeling, and Augmented Reality, as it can provide results with better precision and accuracy. In Medical Imaging, 3D data from Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) has long been used to provide quality visual information for doctors to accurately diagnose different diseases and effectively treat them. Since there are very few labeled medical imaging datasets available, and as it requires a lot of effort from certified professionals, 3D Semi-supervised Learning techniques are important to accurately generate segmentation masks on the medical images. With unlabeled data from CT and MRI modalities, deep learning models that are trained with semi-supervised learning techniques can achieve significantly better performance during evaluation and even achieve results comparable with Fully-Supervised Learning models.

In this report, we implement two 3D Semi-supervised learning frameworks - Cross Pseudo Supervision and Mean Teacher with Uncertainty Masks - to generate 3D segmentation masks on the 3D MR Images of the Left Atrium Dataset from the 2018 Atrial Segmentation Challenge [1] by leveraging unlabeled data. Furthermore, we adopt Lovasz-Softmax Loss [2] to improve the boundary detection and improve the ASD and 95HD scores in the final evaluation.

II. 3D-segmentation V-Net:

Our baseline is adopted from V-Net [3]. VNet is a 3D image segmentation network designed based on a fully convolutional neural network. The VNet network shown in the figure below is the compression path on the left, and the features in the decompress path on the right will be restored to the original size, and padding is used to control the size during the convolution process.

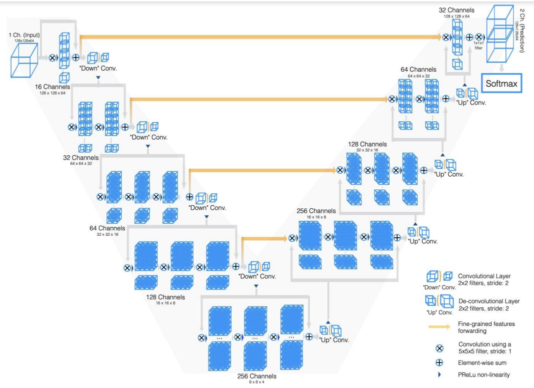


Figure 1.1. Schematic representation of our network architecture. Our custom implementation of Caffe processes 3D data by performing volumetric convolutions. Best viewed in electronic format.

The left side of the VNet network can be seen as different stages. Each stage operates on feature maps of different resolutions by 1-3 convolutional layers. After convolution, the number of feature map channels will increase. The convolution process includes the idea of

[1] Both authors share the same contribution to the report work.
[2] Our code is now available on https://github.com/JR-Guo/ELEC4010N

residual learning. The feature map of the front layer is added to the output of the latter layer after convolution at this stage. The paper points out that the residual link method helps the convergence of the VNet training process.

III. Related Works:

This key objective in this project is to compare two 3D Semi-Supervised Learning Methods. The dataset is very consistent - it only comprises 3D MRIs of the Left Atrium Cavity. Hence, only methods involving ensembling of multiple models were considered as they are known to produce more precise segmentation masks. Several methods were considered, including Cross Pseudo Supervision, FixMatch, FocalMix, and Mean-Teacher model. For the FixMatch method, it involves generating “weakly augmented” and “strongly augmented” images from the original image, which are used to train two different models, and the predictions from the model trained with “weakly augmented” data are used to generate pseudo labels to train the model with “strongly augmented” data. [4] This method is not suitable for this task because the task is a segmentation problem instead of a classification problem. The FocalMix method uses MixUp augmentation for data augmentation on the original image. [5] The augmented images are inputted into parallel networks for detection, after which a bounding box is drawn over any abnormalities in the MRI/CT images through average ensembling. While this method is promising for abnormality detection, it is not appropriate for segmentation tasks. The Mean Teacher model is an improved version of Temporal Ensembling, which performs EMA (exponential moving average) on the model's predicted value. [6] At the same time, Mean Teachers uses EMA for the weights of the student model, as the teacher model is as follows in our experiment. Since this model is effective in generating segmentation masks, we eventually chose to compare the results of the Cross Pseudo Supervision and the Mean Teacher method on this 3D Image Segmentation problem.

IVa. Cross Pseudo Supervision:

In this project, two different semi-supervised learning methods were implemented for the 3D semantic segmentation problem, and their results are compared in Section VI. Cross Pseudo Supervision is a consistency-regularization approach in which two models of the same architecture but different initializations are used. [7] The output predictions from each model is used to generate a one-hot pseudo label map, which supervises the learning of the other model using a cross entropy loss.

Equation 1: $L_S = \left|\frac{1}{D}\right| \sum_{X \epsilon D} \frac{1}{W \times H \times D} \sum_{i=0}^{W \times H \times D} [(l_{CE}(p_{1i}, y_{1i}) + (l_{CE}(p_{1i}, y_{1i})]$

Equation 2: $L_{CPS} = \left|\frac{1}{D}\right| \sum_{X \epsilon D} \frac{1}{W \times H \times D} \sum_{i=0}^{W \times H \times D} [(l_{CE}(p_{1i}, y^*_{2i}) + (l_{CE}(p_{2i}, y^*_{1i})]$

Equation 3: $L = L_S + \lambda L_{CPS}$

As described in Equations 1, 2 and 3, the loss function for this method includes the respective segmentation losses for both models - a standard cross entropy loss between the predictions (p) and their labels (y) - and a weighted (λ) cps loss - the cross entropy loss between a model’s predictions and the pseudo label (y^*) generated from the other model. For the unlabelled data, only a weighted cps loss is used for training.

The 3D V-Net model architecture described in Section III was used for both models. During training, two SGD optimizers were used for both models with a weight decay of 0.0005. Figure 3.1 reports the training of the models.

IVb. Mean Teacher and Uncertainty Mask:

Mean Teacher [6], a popular method which has been proposed in the year 2017 is a method for self-supervised learning in semi-supervised learning. let θ_t equal to time t's parameters in the equation. For the teacher, the model $\theta_t = \alpha\theta_{t-1} + (\alpha - 1)\theta_t$ is used to update the model. Similarly, in the student model, $\Gamma = \Gamma_s + \lambda J(\theta)$ is used to update the model. Since $J(\theta) = \mathbb{E}_{x,\eta,\eta'}[||f(x,\theta',\eta') - f(x,\theta,\eta)||^2]$, Γ_s is classification loss on the student model, λ represents the consistency weight. $J(\theta)$ represents the consistency cost, which is a measure of the distance between the teacher model and the student model's prediction of sample X(here, Mean-Squared Error (MSE) is used to measure, and later experiments try to use KL-divergence to measure).

Mean Teacher performs ensemble on model parameters instead of prediction ensemble. From the EMA formula, it can be understood as a momentum network, which is to replace gradient correlation with model parameter correlation in momentum SGD. Relatively speaking, this ensemble parameter method will make the model more robust, instead of only receiving the update of the current gradient.

Uncertainty Estimation is used to prevent the noisy and unreliable pseudo-label. With the Monte Carlo Dropout, we can estimate the uncertainty in the map. With the guidance of the estimated uncertainty U, we filter out the relatively unreliable (high uncertainty) predictions and select only the certain predictions as targets for the student model to learn from.

$$\mu_c = \frac{1}{T}\sum_t \mathbf{p}_t^c \qquad \text{and} \qquad u = -\sum_c \mu_c \log \mu_c, \qquad \mathcal{L}_c(f', f) = \frac{\sum_v \mathbb{I}(u_v < H)\, \|f'_v - f_v\|^2}{\sum_v \mathbb{I}(u_v < H)},$$

The equation shows how we calculate the **Uncertainty-Aware Consistency Loss** and **Uncertainty Map.** Figure 4.1. shows the loss curve during training in Appendix.

V. Experiments and Results:

We use one NVIDIA Geforce GTX 2080 TI for training, which has a memory size of 11GB. We find out that setting batch size as 4 is the most effective. We also find out that some images slow down the training speed and convergence speed. To solve this, we occasionally take more devices to train parallel. This trick improves the training speed by about 10%. A full-batch training for our total network takes 2.5 hours to 3 hours on GPU for Mean Teacher Uncertainty Mask.

Method & Data	Dice[%]	Jaccard[%]	ASD[voxel]	95HD[voxel]
V-Net & 16-L	78.3	69.0	22.5	9.88
V-Net & 80-L	89.0	80.7	2.07	6.93
CPS & 16L+64UL	86.4	76.6	3.42	12.1

CPS & 32L+48UL *	85.3	74.9	2.32	10.47
MT+UN &16L+64UL	87.2	77.8	2.82	9.31
MT+UN+Lovasz & 16L + 64UL	86.9	76.5	**2.09**	**8.39**
CPS+ALW & 16L + 64UL	**88.4**	**79.5**	2.91	9.56

Table 5. Results from baseline and semi-supervised learning network. 80L means the fully supervised model. **Bold** means the best value within the table. *For reference only

VI. Visualization:

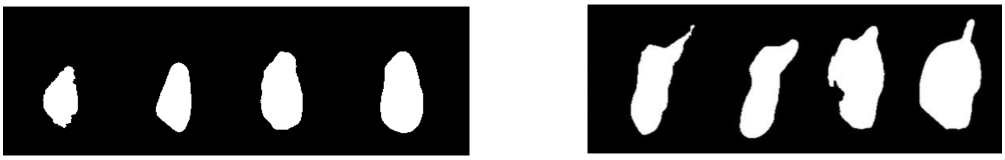


Figure 6.1. Left is 20% labeled data, right is 100% labeled data. Four samples of segmentation are presented. For every sample, our prediction is on the left and the respective ground truth mask is on the right.

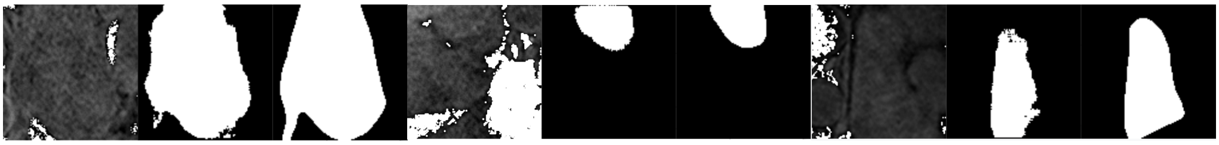
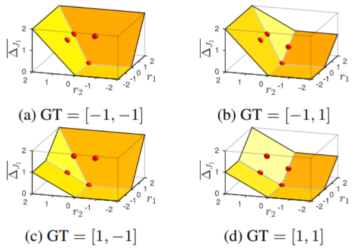


Figure 6.2. Transformed input image is on the left, predicted mask is in the middle, and the segmentation mask is on the right. Three samples of predictions are being demonstrated.

Figure 6.1 demonstrates some visualizations from the Mean-Teacher Method, and figure 6.2 demonstrates some visualizations for the Cross Pseudo Supervision Method. As seen from both methods, the models can predict accurately in the central regions, but the boundaries of the segmentation masks are not accurate and even blurred.

VII. Novelty 1: Lovasz-Softmax Loss

Generally, for 3D medical image segmentation, various losses are used to improve the performance in different evaluation metrics. However, the loss of medical image segmentation can be applied mainly from four different aspects: Distribution-based Loss, Compound Loss, Region-based Loss, and Boundary-based Loss, as shown in Figure 7.1.



(a) GT $= [-1,-1]$

(b) GT $= [-1,1]$

(c) GT $= [1,-1]$

(d) GT $= [1,1]$

Figure 1: Lovász hinge in the case of two pixel predictions for the four possible ground truths GT, as a function of the relative margins $r_i = 1 - F_i(\boldsymbol{x})\, y_i^*$ for $i = 1, 2$. The red dots indicate the values of the discrete Jaccard index.

Figure 7.1. The simple illustration about Lovasz-Softmax Loss.

From the visualization in the above experiment, we figure out that the boundary prediction for segmentation cannot be well organized. So we choose to add Region-based Loss for Lovasz:

- $y_i^* \in \{-1, 1\}$ the ground truth label of pixel i,
- $F_i(\mathbf{x})$ the i-th element of the output scores $\boldsymbol{F}$ of the model, such that the predicted label $\tilde{y}_i = \text{sign}(F_i(\boldsymbol{x}))$,
- $m_i = \max(1 - F_i(\boldsymbol{x})\, y_i^*, 0)$ the hinge loss associated with the prediction of pixel i.

Moreover, by choice of the hinge loss for the vector m, the Lovasz hinge reduces to the standard hinge loss in the case of a single prediction, or when using the Hamming distance instead of the Jaccard loss as a basis for the construction. Figure 7.1. illustrates the extension of the Jaccard loss in the case of the prediction of two pixels, illustrating the convexity and the tightness of the surrogate.

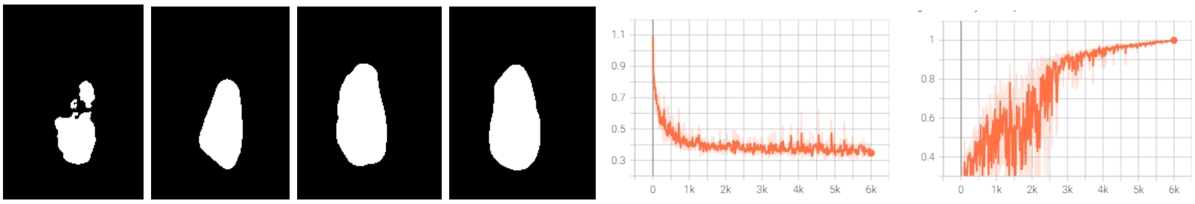


Figure 7.2. Visualization of MT+UN+Lovasz & 16L + 64UL

As seen in Figure 7.2, the predictions are much more smooth with clearer boundaries, but they suffer from uncertainty in the central parts. Section V reports the test accuracies of this novelty.

VIII. Novelty 2: Adaptive Loss Weight for Cross Pseudo Supervision on Unlabelled Data

In Cross Pseudo Supervision, the L_{CPS} acts as a consistency loss between the predictions of two models. However, during training, the accuracy of these models vary significantly. Hence, if the models are inaccurate, using such a consistency loss on unlabelled data may worsen the training of these models (especially when the unlabelled data consists of 80% of the total dataset used). Therefore, this novelty proposes the use of the average dice coefficient computed from the supervised part of the training every epoch to dynamically adjust the weight of this L_{CPS} for unlabelled data. The proposed loss function is as shown in Equation 4.

Equation 4: $L_{CPS} = \overline{Dice}_{Supervised} \times \lambda \times \left|\frac{1}{D}\right| \sum_{X \in D} \frac{1}{W \times H \times D} \sum_{i=0}^{W \times H \times D} [(l_{CE}(p_{1i}, y^*_{2i}) + (l_{CE}(p_{2i}, y^*_{1i})]$

This novelty was implemented with a data mix of 16 Labeled and 64 Unlabeled 3D MRIs, and the result is available in Section V.

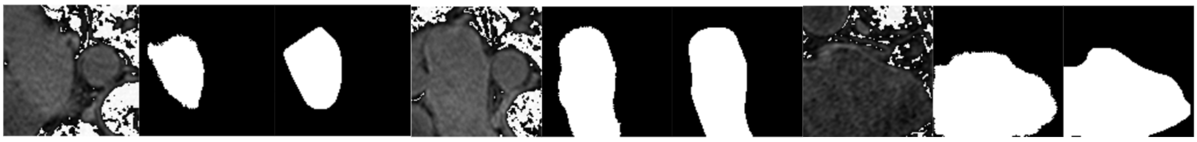


Figure 8.1 Visualization of CPS with Adaptive Loss Weight (ALW) & 16L + 64UL

As seen in the visualizations above, the model is able to accurately predict the central regions of the segmentation masks; however, the predictions at the boundaries can be improved. As seen in the results, Novelty 1 successfully reduces the Average Surface Distance and a combination of the Lovasz-Softmax Loss could be used to further improve the prediction results.

IX. Further Improvements:

There is significant model overfitting because only 16 (or 32) labeled images were used in training. For data augmentation, we used Random Cropping to crop random patches of size (112, 112, 80), Random Rotation, and Random Flipping. More data augmentation techniques such as FixMatch and CutMix could be considered in the future to improve the robustness of our models. According to the original paper on Cross Pseudo Supervision, using CutMix significantly improves the test accuracy especially when the labeled to unlabeled data ratio is small.

Moreover, while the Lovasz-Softmax Loss was implemented in Section VIII to improve the prediction accuracy at the boundaries of the segmentation, this function alone does not suffice because as seen in Section VII, while the segmentation boundaries are smoother, the prediction in the central regions seems to be less accurate compared to the baseline model. Hence, alternate loss functions with different weights could be experimented with to find the optimal loss function that can improve the segmentation results. Finally, a proper hyperparameter sweep could be executed to fine-tune the hyperparameters for better training convergence and higher test accuracy.

X. Conclusion:

This project primarily aims to explore two different 3D Semi-Supervised Learning methods for 3D Segmentation tasks on the MICCAI 2018 Atrial Segmentation Challenge Dataset. The Cross Pseudo Supervision method and the Mean-Teacher method were implemented and trained with different mixes of labeled and unlabeled data, and their test accuracies were compared against baseline models. The test final accuracy of the Mean-Teacher model trained with 20% labeled data was the highest amongst the semi-supervised methods implemented, with a final dice score comparable to that of the baseline model trained with 100% labeled data. From the experiments, we conclude that the Mean-Teacher method may be a more appropriate method to train a model for 3D Segmentation tasks with unlabeled data. However, there is still significant scope to improve our results through data augmentation and experimenting with various loss functions.

Appendix A:

Reference List:

1. The 2018 Atrial Segmentation Dataset
   a. http://atriaseg2018.cardiacatlas.org/
2. The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks
   a. https://arxiv.org/abs/1705.08790
3. ELEC4010N Course Lecture Notes
4. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence
   a. https://arxiv.org/ftp/arxiv/papers/2001/2001.07685.pdf
5. FocalMix: Semi-Supervised Learning for 3D Medical Image Detection
   a. https://arxiv.org/pdf/2003.09108.pdf
6. Uncertainty-aware Self-ensembling Model for Semi-supervised 3D Left Atrium Segmentation
   a. https://arxiv.org/pdf/1907.07034.pdf
7. Semi-Supervised Semantic Segmentation with Cross Pseudo Supervision
   a. https://arxiv.org/pdf/2106.01226.pdf
8. Source for Figure 1 Lovasz-Softmax Loss: https://github.com/JunMa11/SegLoss

Some Additional Figures:

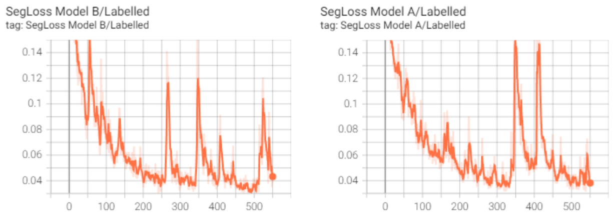


Figure 3.1. From both models, CPS training can get high consistency between two models.

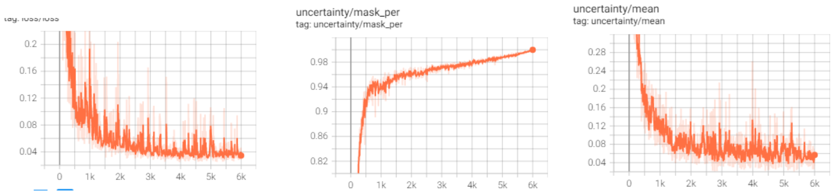


Figure 4.1. The loss curve and uncertainty mask and mean result from MA-UN.

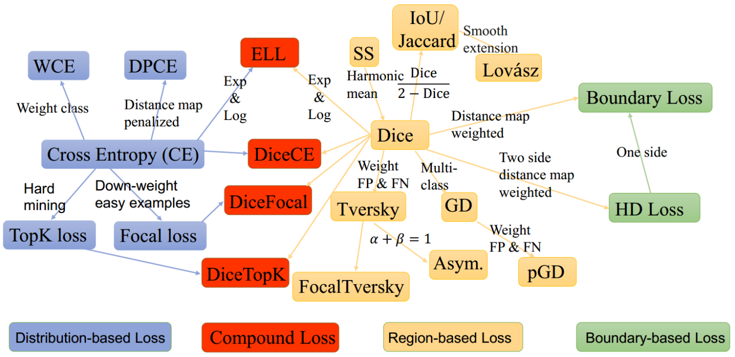


Figure 1 Lovasz-Softmax Loss. The illustration is about 4 kinds of medical image segmentation loss.