Semi-supervised Augmented 3D-CNN for FLARE22 Challenge

Zining Chen[1], Tianyi Wang[1], Shihao Han[1], Yinan Song[1], Shichao Li[1]

Beijing University of Posts and Telecommunications, Beijing, China
chenzn@bupt.edu.cn

Abstract. Abdominal organ segmentation has been used in many important clinical applications, however, cases with accurate labels require huge manual labour and financial resources. As a potential alternative, semi-supervised learning can explore useful information from unlabeled cases, with only few labeled cases involved. Therefore, we propose our baseline model using augmented 3D-UNet and adopt semi-supervised method–Mean Teacher, to make quantitative evaluation on the FLARE2022 validation cases. Our method achieves average dice similarity coefficient (DSC) of 62.16%, Normalized Surface Distance (NSD) of 62.27%, running time of 9.58s, and AUC of GPU and CPU is only 7424 and 199 respectively, which surpasses almost all other teams on resource consumption, demonstrating the effectiveness of our methods.

Keywords: Abdominal organ segmentation · Semi-supervised learning

1 Introduction

Computed Tomography (CT) has long been regarded as an effective therapeutic method in clinical workflow and is capable of improving patient treatment by visualization of abnormal organs. With the rapid development of deep learning, semantic segmentation in medical image plays an important role in clinical practice and is used in radiotherapy to accurately delineate tumors and treat certain cancers [9].

However, huge amount of labeled medical image are required for fully-supervised segmentors, which is not only laborious and time-consuming, but also cost-intensive and conse-quently. Thus, semi-supervised learning is proposed to explore useful information from unlabeled cases, which is a combination of supervised learning and unsupervised learning. The basic process uses the existing labeled cases to pseudolabel the remaining unlabeled data, so as to effectively help increase the information in training data, which can strengthen the consistency of the prediction of unlabeled data and labeled data through the regularization in loss function. For semantic segmentation, convolutional neural network(CNN) has achieved a dominant position in the field of medical image segmentation, especially Unet and its various modified versions by adjusting the network structure, e.t.c. adding various attention mechanisms and feature fusion structures, aiming to fit a more powerful model on the limited data.

However, the performance of semi-supervised learning methods in medical image segmentation is still limited, most of which are only able to process 2D images while FLARE2022 challenges focus on 3D volumes in clinical practice. Difficulties mainly stem from four aspects: 1) Variations in field-of-views, shape and size of different organs. 2) Difficulty in using unlabeled data. 3) Diversity of data source in term of multi-center, multi-phase and multi-vendor cases. 4) Limited GPU memory size and high computation.

In this stage, we develop a semi-supervised baseline composed of backbone network 3D-UNet[5] and semi-supervised method Mean Teacher [11] to effectively and efficiently tackle FLARE2022 challenges. The model aims to obtain the rough location of target organs from the whole CT volume. In this way, the background can be preferentially screened, which is more conducive to the identification of target organs. To overcome Temporal Ensembling, a common semi-supervised potential problem, we use Mean Teacher method which effectively update teacher model weights instead of hard label predictions.

2 Method

This whole-volume-based semi-supervised segmentation framework is composed of backbones network 3D-UNet and semi-supervised method Mean Teacher. A detail description of the method is as follows.

2.1 Preprocessing

The baseline method includes the following preprocessing steps:

- Delete abnormal data.
- Reorientation image to target direction.
- Resample image to fixed size: [160, 160, 160].
- Intensity normalization: Apply a z-score normalization based on the mean and standard deviation of the intensity values.
- Clip image in range of [-600, 600].
- Convert mask labels into one-hot coding formation.

2.2 Proposed Method

As mentioned in Figure 1 and Figure 2, our framework uses Mean Teacher method on semi-supervised learning, and apply a two-stage coarse-and-fine segmentation method based on 3D-UNet to extract abdominal features. A detail description of our method is as follows.

First, based on the winning solution in FLARE 2021, we accordingly adopt the applied 3D-UNet, as illustrated in Figure 1. The proposed backbone can learn from sparse annotations and provides a dense 3D segmentation mask corresponding to the 3D image, showing great robustness on various abdominal organ segmentation tasks.

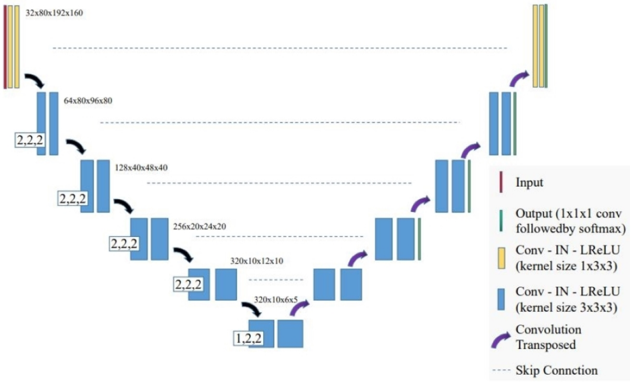

Fig. 1. 3D U-Net

Second, method used on semi-supervised learning is Mean Teacher, illustrated in Figure 2, which is composed of a student and teacher network. Input a batch of labeled and unlabeled cases with random noise to student network, and calculate the supervised loss on labeled cases to update student network, after which the teacher network weights are updated as an exponential moving average of the student network weights. Also, input batch to teacher and student network to calculate the comparative loss on unlabeled cases, then two losses are summed as the final loss for gradient descent to update student network.

Third, to improve inference speed and reduce resource consumption, we recommend using ONNX or TensorRT to speed up the inference process.

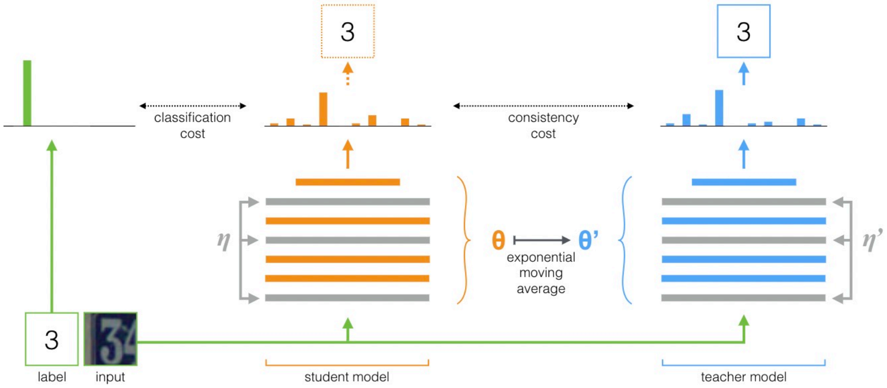

Fig. 2. Mean Teacher Method

2.3 Post-processing

We tend to use a connected component analysis of segmentation mask applied on model output in next stage.

3 Experiments

3.1 Dataset and evaluation measures

The FLARE2022 dataset is curated from more than 20 medical groups under the license permission, including MSD [9], KiTS [3,4], AbdomenCT-1K [8], and TCIA [2]. The training set includes 50 labelled CT scans with pancreas disease and 2000 unlabelled CT scans with liver, kidney, spleen, or pancreas diseases. The validation set includes 50 CT scans with liver, kidney, spleen, or pancreas diseases. The testing set includes 200 CT scans where 100 cases has liver, kidney, spleen, or pancreas diseases and the other 100 cases has uterine corpus endometrial, urothelial bladder, stomach, sarcomas, or ovarian diseases. All the CT scans only have image information and the center information is not available.

The evaluation measures consist of two accuracy measures: Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD), and three running efficiency measures: running time, area under GPU memory-time curve, and area under CPU utilization-time curve. All measures will be used to compute the ranking. Moreover, the GPU memory consumption has a 2 GB tolerance.

3.2 Implementation details

Environment settings The environments and requirements are presented in Table 1.

Windows/Ubuntu version	Ubuntu 18.04.5 LTS
CPU	Intel(R) Xeon(R) Gold 5218R CPU@2.10GHz
RAM	32×4GB; 3200MT/s
GPU (number and type)	Two Nvidia Tesla T4 16G
CUDA version	10.2
Programming language	Python 3.7
Deep learning framework	Pytorch (Torch 1.8.0, torchvision 0.9.0)

Table 1. Environments and requirements.

Training protocols Training protocols are listed in Table 2. First, due to limited GPU resources and speed requirements, we set patch size as [160,160,160]. Furthermore, imbalance on field-of-views, shape and size of different organs further render difficulty on hard samples. Therefore, we design DiceLoss-based MultiDiceLoss, utilizing segmentation mask distribution to caluculate weights on different organs. Also, two sets of indices are designed to be iterated in single iteration, during which we sample from both the primary indices and secondary indices, and labeled cases are iterated as many times as needed. Finally, we use loss value and visulization on training and validation set to select optimal model hyperparameters. Visulization is shown in Figure 3 and 4.

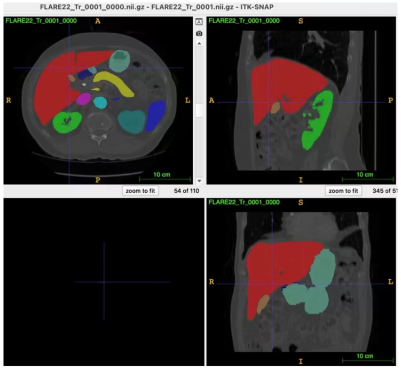
FLARE22_Tr_0001_0000.nii.gz - FLARE22_Tr_0001.nii.gz - ITK-SNAP
FLARE22_Tr_0001_0000
A
R
L
P
10 cm
zoom to fit
54 of 110
S
I
345 of 5

Fig. 3. Test on Train Set

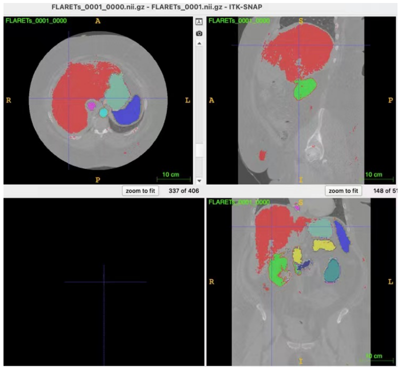

Fig. 4. Test on Validation Set

Network initialization	"he" normal initialization
Batch size	4
Patch size	160×160×160
Loss	MultiDiceLoss and FocalLoss [7]
Total iterations	8000
Optimizer	AdamW with weight decay =1e-4 and $\beta = 0.9$
Initial learning rate (lr)	0.002
Training time	15.5 hours
Number of model parameters	1.50M[1]
Number of flops	73.09G[2]

Table 2. Training protocols.

4 Results and discussion

Unlabeled cases enrich image information and effectively reduce overfitting. Ablation study is used to verify the necessity of unlabeled cases. According to visualization in Figure 3, if labeled images are only input to the whole network, loss decreases fast but prones to overfitting. The use of unlabeled cases can effectively tackle the overfitting problem with more feature information, meanwhile pseudolabel can be obtained through training.

To solve the above problems, method used on semi-supervised learning–Mean Teacher works well. It is composed of a student and teacher network. The whole process includes, input a batch of labeled and unlabeled cases with random noise to student network, and calculate the supervised loss on labeled cases to update student network, after which the teacher network weights are updated as an exponential moving average of the student network weights. Then, input batch to teacher and student network to calculate the comparative loss on unlabeled cases, and two losses are summed as the final loss for gradient descent to update student network.

Our mean DSC scores is relatively low, especially on abnormal cases and lesion-affected organs, e.t.c. RAG, LAG, Gallbladder. Reasons can be categorized into insufficient training, ordinary one-stage coarse framework, few processing strategies. Therefore, in next stage, we will make improvements on accurate extraction of image features by desiging two-stage coarse-and-fine framework. Also, we will use other semi-supervised methods, e.t.c pseudolabel on unlabeled cases [1], data augmentation method [10]. Last but not least, we will try different training strategies and preprocessing/postprocessing methods.

4.1 Quantitative results on validation set

Table 3 illustrates the results on validation set. Only using labeled data results in quick overfitting. With the design of Mean Teacher method, our model can effectively alleviate this problem.

4.2 Segmentation efficiency results

The running time is 9.58s and maximum used GPU memory is 2011 MB. To accelerate inference process, we tend to use ONNX or TensorRT. To further decrease the use of GPU memory, we attempt to design lighter network architechure and other data processing methods. Finally, our method achieves average dice similarity coefficient (DSC) of 62.16%, Normalized Surface Distance (NSD) of 62.27%, running time of 9.58s, and AUC of GPU and CPU is only 7424 and 199 respectively.

5 Conclusion

Our method performs high generalization and robustness on most organs, such as liver, kidney and spleen in terms of DSC scores. Also, our AUC of GPU

Mean DSC	0.60
Liver	0.84
RK	0.72
Spleen	0.74
Pancreas	0.48
Aorta	0.81
IVC	0.69
RAG	0.44
LAG	0.40
Gallbladder	0.22
Esophagus	0.60
Stomach	0.69
Duodenum	0.44
LK	0.75

Table 3. Quantitative results.

and CPU is only 7424 and 199 respectively, which surpasses almost all other teams on resource consumption, demonstrating the effectiveness of our methods. However, RAG, LAG and Gallbladder performs relatively bad as a result of the inter-patient and anatomical variability of volume and shape. Meanwhile, lesion-affected organ is also a critical reason for the poor segmentation performance. Therefore, we consider investigations on fine-stage segmentation network, e.t.c nn-UNet [6], preprocessing methods on certain organs, and postprocessing strategies for future work, to obtain a more accurate boundary segmentation to increase DSC scores.

Acknowledgements The authors of this paper declare that the segmentation method they implemented for participation in the FLARE 2022 challenge has not used any pre-trained models nor additional datasets other than those provided by the organizers.

References

1. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision (2021) 6
2. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al.: The cancer imaging archive (tcia): maintaining and operating a public information repository. Journal of Digital Imaging **26**(6), 1045–1057 (2013) 4
3. Heller, N., Isensee, F., Maier-Hein, K.H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., et al.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. Medical Image Analysis **67**, 101821 (2021) 4

4. Heller, N., McSweeney, S., Peterson, M.T., Peterson, S., Rickman, J., Stai, B., Tejpaul, R., Oestreich, M., Blake, P., Rosenberg, J., et al.: An international challenge to use artificial intelligence to define the state-of-the-art in kidney and kidney tumor segmentation in ct imaging. American Society of Clinical Oncology **38**(6), 626–626 (2020) 4
5. iek, zgün, A.A.L.S.S.: 3d u-net: Learning dense volumetric segmentation from sparse annotation. International conference on medical image computing and computer-assisted intervention pp. 424–432 (2016) 2
6. 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) 7
7. Lin T Y, Goyal P, G.R.: Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision pp. 2980–2988 (2017) 5
8. Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., Cao, S., Zhang, Q., Liu, S., Wang, Y., Li, Y., He, J., Yang, X.: Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6695–6714 (2022) 4
9. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019) 1, 4
10. Souly, N., Spampinato, C., Shah, M.: Semi supervised semantic segmentation using generative adversarial network. In: 2017 IEEE International Conference on Computer Vision (ICCV) (2017) 6
11. Tarvainen A, V.H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems (2017) 2