

# Semi-supervised Context-aware Multi-Organ Segmentation

Jiacheng Wang<sup>1</sup> and Yongkang Ma<sup>1</sup>

Manteia Technologies Co.,Ltd, Xiamen, China  
{wangjiacheng, yongkangma}@manteiatech.com

**Abstract.** The automatic segmentation of organs at risk is extremely dedicated to the clinical assistance that can significantly reduce the clinical resource cost. However, training such a good enough model usually requires a large amount of labeled data, or the model performance is likely to meet heavy drop. Semi-supervised training strategies are proved to be an effective solution to reduce the reliance of labeled data. In this paper, we develop a powerful semi-supervised learning framework to address the label-efficient multi-organ segmentation. The experiments are conducted on the MICCAI FLARE 2022 dataset, where the results show that the semi-supervised learning strategy has significant performance boost.

**Keywords:** Semi-supervised, Multi-Organ segmentation

## 1 Introduction

The automatic segmentation of organs at risk [5,6,12,10,25,14,13,22] is extremely dedicated to the clinical assistance that can significantly reduce the clinical resource cost. However, training such a good enough model usually requires a large amount of labeled data, or the model performance is likely to meet a heavy drop. Particularly for the clinics, it is well-known that the dense-level annotation needs expertise knowledge to make the precise decision. Thus, simply involving a large group of annotators to handle the annotation task is not helpful. To develop a label-efficient model seems to be the only one alternation that is also desperately needed nowadays.

There are plenty of wise solutions to reduce the reliance of labeled data, such as weakly supervised learning, self-supervised learning during model pretraining [24,21,11,2,1,20], and the most popular one of which is the semi-supervised learning methods [19,7,23,4,18]. As a famous research direction, a lot of studies have been conducted in distinct areas and imaging modalities, such as natural images, radiation images, fundus iamges, and so on. No matter what the imaging is, these methods have similar underlying principles that are to produce convincing pseudo labels for semi-supervision. In this group, the most simple and straightforward tool is to produce pseudo label using trained models and then using a post-processing strategy to reduce the false positives. Here, in this work, we follow this group and apply it into the 3D context-aware organ segmentation models.

## 2 Method

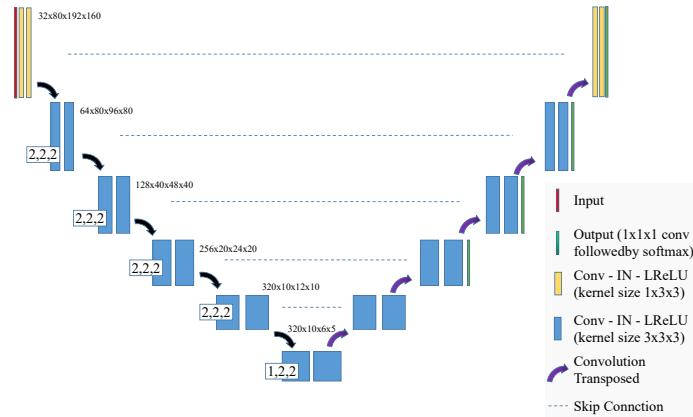
### 2.1 Preprocessing

The preprocessing steps include 1) reorienting images to the left-posterior-inferior (LPI) view by flipping and reordering, 2) resampling images to keep a fixed size (160, 160, 160), and 3) normalizing images to reduce the distribution error by z-score normalization.

### 2.2 Proposed Method

Our model follows the standard 3D U-Net design to achieve the coarse-to-fine organ segmentation. It consists of two stages, where the first stage is to coarsely localize the organs and the second stage is to precisely segment the boundaries.

Specifically speaking, the two stages share same network architecture design as Fig. 1 shows, which is just a standard 3D U-Net. They differ in the inputting data where the first stage takes the original data as input while the second stage takes the cropped prediction of the first stage as input.



**Fig. 1.** The used network architecture of one stage.

### 2.3 Unlabel Data

To use the unlabeled volumes, we first train a good segmentation model under full supervision using the labeled data. Then, the model calculates the prediction of each unlabeled volume to obtain the pseudo label. As the existing false positives heavily affect the label quality, we apply the post-processing strategy like connective field analysis to improve the predictions. After generating the pseudo labels, we merge these data into the label data and re-train the segmentation models from scratch.

## 2.4 Objectives

We use the summation between Dice loss and cross entropy loss because compound loss functions have been proved to be robust in various medical image segmentation tasks [15].

## 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 [17], KiTS [8,9], AbdomenCT-1K [16], and TCIA [3]. 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 detailed environment settings have been shown in Table 1 where we use a single Titan 3090 GPU with 24 GB memory to train the entire network.

**Table 1.** Development environments and requirements.

|                         |   |
|-------------------------|---|
| Windows/Ubuntu version  | Ubuntu 16.04.5 LTS                      |
| CPU                     | Intel(R) Core(TM) i9-7700X CPU@3.30GHz  |
| RAM                     | 16×4GB; 2.67MT/s                        |
| GPU (number and type)   | One NVIDIA 3090 24G                     |
| CUDA version            | 11.0                                    |
| Programming language    | Python 3.9                              |
| Deep learning framework | Pytorch (Torch 1.10, torchvision 0.2.2) |
| Specific dependencies   |   |
| (Optional) Link to code |   |

**Training protocols** The details of training protocols have been shown in Table 2 and Table 3. We use the standard “he” normal initialization to initialize the model parameters with the consideration to avoid overfitting. The batch size is set to four considering the memory limitation and the computation efficiency. For the optimizer, we choose to utilize the AdamW optimizer empirically with an initial learning rate of 0.001. For stable training process, we apply the decay schedule to reduce the learning rate half each 200 epochs.

**Table 2.** Training protocols at the first stage.

|                            |                             |
|----------------------------|-----------------------------|
| Network initialization     | “he” normal initialization  |
| Batch size                 | 4                           |
| Patch size                 | $192 \times 192 \times 192$ |
| Total epochs               | 1000                        |
| Optimizer                  | AdamW                       |
| Initial learning rate (lr) | 0.001                       |
| Lr decay schedule          | halved by 200 epochs        |
| Training time              | 72.5 hours                  |
| Number of model parameters | 41.22M <sup>1</sup>         |
| Number of flops            | 59.32G <sup>2</sup>         |
| CO <sub>2</sub> eq         | 1 Kg <sup>3</sup>           |

**Table 3.** Training protocols at the second stage.

|                            |                             |
|----------------------------|-----------------------------|
| Network initialization     | “he” normal initialization  |
| Batch size                 | 4                           |
| Patch size                 | $160 \times 160 \times 160$ |
| Total epochs               | 1000                        |
| Optimizer                  | AdamW                       |
| Initial learning rate (lr) | 0.001                       |
| Lr decay schedule          | halved by 200 epochs        |
| Training time              | 72.5 hours                  |
| Number of model parameters | 41.22M <sup>4</sup>         |
| Number of flops            | 59.32G <sup>5</sup>         |
| CO <sub>2</sub> eq         | 1 Kg <sup>6</sup>           |

## 4 Results and discussion

### 4.1 Quantitative results on validation set

The averaged DSC scores of models trained under full-supervision and semi-supervision on the 10 cases validation set are respectively 0.8345 and 0.8523. It could be seen that the pseudo labeled data has performance boost for the segmentation modeling.

### 4.2 Qualitative results on validation set

This part is optional during validation phase since you do not have validation ground truth.

**Table 4.** Test results.

| Organs      | DSC score |
|-------------|-----------|
| Liver       | 0.9410    |
| RK          | 0.9003    |
| Spleen      | 0.9005    |
| Pancreas    | 0.7585    |
| Aorta       | 0.9352    |
| IVC         | 0.8513    |
| RAG         | 0.6579    |
| LAG         | 0.7139    |
| Gallbladder | 0.6854    |
| Esophagus   | 0.7522    |
| Stomach     | 0.8021    |
| Duodenum    | 0.6951    |
| LK          | 0.8823    |
| Mean        | 0.8058    |

### 4.3 Limitation and future work

The first problem of semi-supervised organ 3D segmentation is the memory limitation. As the previous tables have shown, the batch size is limited to 4 only which is a relatively small number. When performing the semi-supervised learning process, the data in one minibatch requires the strongly correct data to guide the optimization. With a limited and small mini-batch, the pseudo data will cover a large proportion so that the optimization will be affected and the final performance is poor.

## 5 Conclusion

In this work, we take the attempt to develop a semi-supervised learning framework to reduce the label reliance for multi-organ segmentation. We build a coarse-to-fine 3D organ segmentation model and train it under a supervised training manner. The model after training on the unlabeled data has a slight performance improvement. Due to the memory limitation, the performance has not been improved a lot, which can be further discussed in the future work.

**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. The proposed solution is fully automatic without any manual intervention.

## References

1. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 132–149 (2018) [1](#)
2. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. arXiv preprint arXiv:2006.09882 (2020) [1](#)
3. 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) [3](#)
4. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 113–123 (2019) [1](#)
5. Day, E., Betler, J., Parda, D., Reitz, B., Kirichenko, A., Mohammadi, S., Miften, M.: A region growing method for tumor volume segmentation on pet images for rectal and anal cancer patients. *Medical physics* **36**(10), 4349–4358 (2009) [1](#)
6. Geets, X., Lee, J.A., Bol, A., Lonneux, M., Grégoire, V.: A gradient-based method for segmenting fdg-pet images: methodology and validation. *European journal of nuclear medicine and molecular imaging* **34**(9), 1427–1438 (2007) [1](#)

7. Grill, J.B., Strub, F., Althé, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent—a new approach to self-supervised learning. *Advances in neural information processing systems* **33**, 21271–21284 (2020) [1](#)
8. 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) [3](#)
9. 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) [3](#)
10. Ibragimov, B., Xing, L.: Segmentation of organs-at-risks in head and neck ct images using convolutional neural networks. *Medical physics* **44**(2), 547–557 (2017) [1](#)
11. Jaus, A., Yang, K., Stiefelhagen, R.: Panoramic panoptic segmentation: Towards complete surrounding understanding via unsupervised contrastive learning. *arXiv preprint arXiv:2103.00868* (2021) [1](#)
12. Kerhet, A., Small, C., Quon, H., Riauka, T., Schrader, L., Greiner, R., Yee, D., McEwan, A., Roa, W.: Application of machine learning methodology for pet-based definition of lung cancer. *Current oncology* **17**(1), 41 (2010) [1](#)
13. Li, S., Xiao, J., He, L., Peng, X., Yuan, X.: The tumor target segmentation of nasopharyngeal cancer in ct images based on deep learning methods. *Technology in Cancer Research Treatment* **18** (2019) [1](#)
14. Liu, Z., Liu, X., Xiao, B., Wang, S., Miao, Z., Sun, Y., Zhang, F.: Segmentation of organs-at-risk in cervical cancer ct images with a convolutional neural network. *Physica Medica* **69**, 184–191 (2020) [1](#)
15. Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X., Martel, A.L.: Loss odyssey in medical image segmentation. *Medical Image Analysis* **71**, 102035 (2021) [3](#)
16. 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* (2021). <https://doi.org/10.1109/TPAMI.2021.3100536> [3](#)
17. 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) [3](#)
18. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems* **33**, 596–608 (2020) [1](#)
19. 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) [1](#)
20. Tian, Y., Henaff, O.J., van den Oord, A.: Divide and contrast: Self-supervised learning from uncurated data (2021) [1](#)
21. Wang, W., Zhou, T., Yu, F., Dai, J., Konukoglu, E., Van Gool, L.: Exploring cross-image pixel contrast for semantic segmentation. *arXiv preprint arXiv:2101.11939* (2021) [1](#)

22. Wang, X., Yang, G., Zhang, Y., Zhu, L., Dai, Z.: Automated delineation of nasopharynx gross tumor volume for nasopharyngeal carcinoma by plain ct combining contrast-enhanced ct using deep learning. *Journal of Radiation Research and Applied Sciences* **13**(1), 568–577 (2020) [1](#)
23. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 10687–10698 (2020) [1](#)
24. Zhao, X., Vemulapalli, R., Mansfield, P., Gong, B., Green, B., Shapira, L., Wu, Y.: Contrastive learning for label-efficient semantic segmentation. *arXiv preprint arXiv:2012.06985* (2020) [1](#)
25. Zhu, J., Zhang, J., Qiu, B., Liu, Y., Liu, X., Chen, L.: Comparison of the automatic segmentation of multiple organs at risk in ct images of lung cancer between deep convolutional neural network-based and atlas-based techniques. *Acta Oncologica* **58**(2), 257–264 (2019) [1](#)