Selected Partially Labeled Learning for Abdominal Organ and Pan-Cancer Segmentation

Yuntao Zhu(✉), Liwen Zou, Linyao Li, and Pengxu Wen

Department of Mathematics, Nanjing University, Nanjing, China
YuntaoZhu7@smail.nju.edu.cn

Abstract. Obtaining labeled data from medical images is very expensive and labor intensive. At the same time, the large number of existing publicly available medical image datasets are usually labeled with only some of the organs as target regions, while other organs in the image are ignored. It is a challenge to train a neural network to segment all labeled categories using only partially labeled data. We design a compound loss, the selected partially cross entropy and dice loss, that allows the neural network to learn specific categories from partially labeled data. In addition, we improve the inference and training process of nnU-Net to reduce computational resources and accelerate inference. Experiments demonstrate that our method achieves the average Dice Similarity Coefficient of 0.8514 and 0.1514 on 13 abdominal organ and tumor segmentation tasks, and enables the network to efficiently segment specific categories from partially labeled data. Moreover, it significantly improves the inference speed, with an average running time of 21.8 s, and uses only an average of 2531 MB of maximum GPU memory.

Keywords: Partially labeled learning · Accelerate inference · Lightweight network

1 Introduction

Medical image segmentation aims to extract and quantify regions of interest in biological tissue or organ images. The results of target organ segmentation have many important clinical applications, such as organ quantification, surgical planning, and disease diagnosis. In recent years, deep learning-based methods have been widely used to automatically segment abdominal organs. Among these methods, nnU-Net [11] is a popular and robust framework that has won a number of organ segmentation challenges. Although it is convenient for fully supervised organ segmentation tasks and provides a solid baseline result by automatically setting network hyperparameters, this approach does not support weakly supervised segmentation and the inference process is computationally expensive and time consuming. Numerous studies have shown that the methodological performance of deep neural networks often relies heavily on the availability of large,

J. Ma and B. Wang (Eds.): FLARE 2023, LNCS 14544, pp. 209–221, 2024.
https://doi.org/10.1007/978-3-031-58776-4_17

high-quality labeled datasets for organ segmentation tasks. In order to learn robust data representations for robust and efficient medical image segmentation, we need large datasets with thousands of labeled or unlabeled data for supervised, weakly supervised, and self-supervised learning. But, the annotation of 3D medical images is a difficult and laborious task. Thus, depending on the task, only a bare minimum of images and target structures is usually annotated. This results in a situation where a zoo of partially labeled datasets is available to the community. In this context, the organizer of FLARE2023 build a large-scale and diverse abdomen CT dataset, including 4000 CT scans from servel medical datasets. There are 2200 labeled data and 1800 unlabeled data available. Compared with FLARE 2021–2022 [17,18], the challenge for FLARE 2023 is how to leverage the large amount of partial labels and unlabeled data to improve the segmentation performance while taking into account efficient inference.

In recent years, there has been a rapid evolution of semi-supervised and self-supervised learning methods [24,31]. These techniques typically learn better representations by utilizing unlabeled data, ultimately improving segmentation performance. On the one hand, one frequently employed approach in semi-supervised learning is pseudo-labeling. This method pairs the segmentation results of the network on unlabeled data as pseudo-labels, adds them to the training set, and repeats the process over several iterations. On the other hand, integrating potentially valuable additional information from different datasets, which are partially labeled, can provide more information about different anatomical target structures or related details, as well as different types of pathology. Therefore, recent advances in weak supervision explore how partially annotated datasets can train a model to segment all annotated categories [12]. Early methods considered unlabeled organs as background [4,21] and imposed penalties for overlapping predictions based on mutual exclusivity of organs [5,22]. [26] transforms the cross-entropy loss and dice loss by assigning unlabeled data from partially labeled data to the background class. [3,13,30] predict just one structure of concern per forward pass through the integration of category information at various network stages. [14] use of partial cross entropy and intraclass gray regular terms allows segmentation under weak supervision. [25] ignores the channels where unlabeled categories are located, designs a loss function that mixes binary cross-entropy and dice loss, and can handle the task of category overlap in partial labeling learning. However, there is a lack of methods that utilize both pseudo-labeled data and partially labeled learning techniques to handle organ, tumor segmentation tasks like FLARE23 that contain partially labeled and unlabeled data.

In this paper, we present a framework that utilizes both pseudo-labeled and partially labeled learning by designing a selected partially loss. We also improve nnU-Net for efficient inference and less computational resource respectively. Specifically, we choose to merge 13 organ classes of pseudo-labels and partial labels, while leaving the remaining classes unchanged, resulting in a partial labeling of the tumor. The selected partially loss, which is a combination of cross-entropy loss and dice loss, introduces a selected class mask to deter-

mine whether the class loss will compute and backward gradient. Otherwise, we find that the resampling process in the inference is time-consuming. To address this issue, we have rewritten the implementation of the resampling method and utilized a smaller network and lower resolution to minimize the computational requirements during inference.

Our main contributions are summarized as follows:

- We present a new approach, selected partially loss, which enables the use of both pseudo label and partial label data, thereby expanding the potential applications of current segmentation models.
- We optimize the time-consuming components of the resampling code in nnU-Net.
- The experiment shows that our method improves the detectability of the segmentation network for the selected class. This outperforms the baseline by 5% points for the Dice Similarity Coefficient (DSC).

2 Method

2.1 Preprocessing

For image prepossessing, all of our settings follow the default nnU-NetV2.

- Statistical analysis is conducted on data pertaining to volume spacing and foreground intensity.
- CT images are clipped at the 0.5 and 99.5 percentiles of foreground voxels.
- All images are normalized through the subtraction of the mean and division by the standard deviation.
- The volume is then resampled to a target spacing of (2.42,1.95,1.95).

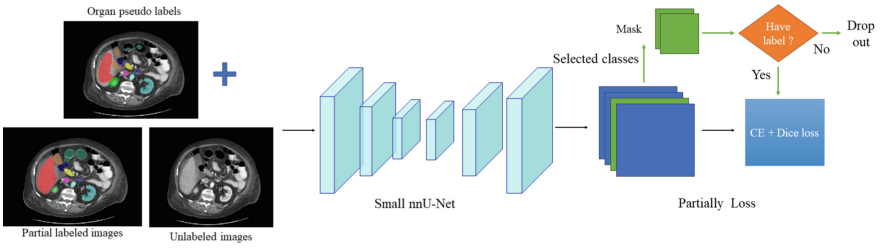

Fig. 1. Overview of our framework. Our framework consists of three parts. Firstly, we construct a training set by combining pseudo-labels. Secondly, we reduce computation costs by using a small nnU-Net. Lastly, we train nnU-Net by a selected-partially-loss so that it can learn from both unlabeled and partially-labeled data.

2.2 Proposed Method

In Fig. 1 we present an overview of our framework, which consists of three components. We filter the data by pseudo and select 300 cases as the training set. We then train a small nnU-Net using a compound partially loss on lower resolution. And our compound partially loss main refer to [14,25,26].

Fusion of Pseudo-labels and Partial-Labels. We use two pseudo-labels generated by [10,27], consisting of 13 organ categories for all 4000 cases. First, We calculate the DSC of the two pseudo-labels, evaluate their differences, and filter out the samples with DSC greater than 0.85. we sort them by their ID numbers. Subsequently, we select first 200 cases from partially labeled CT volumes and first 100 cases from unlabeled CT volumes to construct the training set. Then, the pseudo-labels are merged with the selected cases that do not contain the ground truth annotation of the class. Therefore, for the 300 cases, there are 13 organ labels (ground truth or pseudo) and tumor is partially labeled. All of our results use the pseudo-labels generated by the two FLARE 2022 methods.

Problem Definition. We begin with a dataset D, with N image and label pairs $D = \{(x, y)_1, ..., (x, y)_N\}$. In the dataset, every image voxel $x_i, i \in [1, I]$, is assigned to one class $c \in C$, where C is the label set associated to dataset D. Since the tumor is included in some organs commonly, but the pseudo label does not annotate the tumor. This implies that the network must predict multiple classes for one voxel to account for the inconsistent class definitions. To resolve the issue of label inconsistency, we separate the segmentation results for each class by applying a sigmoid activation function to replace the softmax activation function on the dataset.

Partially Loss for Selected Categories. We employ the binary cross-entropy (BCE) loss and the dice loss for each class over all $B, b \in [1, B]$, images in a batch:

$$L_c = \frac{1}{B \times I} \sum_{b,i} BCE(\hat{y}_{i,b,c},\ y_{i,b,c}) - \frac{2 \sum_{b,i} \hat{y}_{i,b,c}\ y_{i,b,c}}{\sum_{b,i} \hat{y}_{i,b,c} + \sum_{b,i} y_{i,b,c}} \tag{1}$$

We modify the loss function to be calculated only for classes that are annotated in the corresponding partially labeled dataset [4,21].

This partially loss formalize as follow:

$$L = \frac{1}{\sum_{b,c} \mathbb{I}_{b,c}^{(h)}} \sum_{b,c} \Big(\frac{\mathbb{I}_{b,c}^{(h)}}{I} \sum_{i} BCE(\hat{y}_{i,b,c},\ y_{i,b,c}) - \frac{2 \sum_{i} \mathbb{I}_{b,c}^{(h)}\ \hat{y}_{i,b,c}\ y_{i,b,c}}{\sum_{i} \mathbb{I}_{b,c}^{(h)}\ \hat{y}_{i,b,c} + \sum_{i} \mathbb{I}_{b,c}^{(h)}\ y_{i,b,c}} \Big) \tag{2}$$

$$\mathbb{I}_{b,c}^{(h)} = \begin{cases} 0,\ if\ c \in S \text{ and } h = False, \\ 1,\ otherwise, \end{cases}$$

where $c \in S$ is the selected class set, we set $S = \{tumor\}$, h is false if the ground truth data does not include the class c, otherwise it is true. The loss use the summation between dice loss and binary cross entropy loss because compound loss functions have been proved to be robust in various medical image segmentation tasks [15].

Table 1. Network architecture and inference process.

Channels in the first stage	16
Convolution number per stage	2
Patch size	128 × 128 × 128
Downsampling times	4
inference process	(Sigmoid, Threshold, Resample)
Deep supervision	True

Speeding Inference. In order to improve inference speed and reduce resource consumption, we use a small-size network structure in reference [10]. And we change the default resampling function and order, which effectively speeds up the inference. The setup of network architecture and inference process are presented in Table 1. Comparison of different strategy settings in Table 2 . The default is full resolution setting of nnU-Net and the small is low resolution modified. The tiny is the first stage of the cascade network that we design to have a lower resolution. However, we do not use the cascade network as the final docker submission because it does not improve the accuracy and speed of the segmentation results.

Table 2. Comparison of different strategy settings. The order of axes of input patch size and spacing is (z,y,x).

Settings	Default	Small	Tiny
Channels in the first stage	32	16	8
Convolution number per stage	2	2	2
Patch size	56 × 192 × 160	128 × 128 × 128	80 × 96 × 96
Downsampling times	5	4	4
Input spacing	(2.5, 0.8, 0.8)	(2.42, 1.95, 1.95)	(5, 3.9, 3.9)

2.3 Post-processing

We do not perform any post-processing, such as connected component analysis or testing time augmentation, during our inference.

3 Experiments

3.1 Dataset and Evaluation Measures

The FLARE 2023 challenge is an extension of the FLARE 2021–2022 [17] [18], aiming to aim to promote the development of foundation models in abdominal disease analysis. The segmentation targets cover 13 organs and various abdominal lesions. The training dataset is curated from more than 30 medical centers under the license permission, including TCIA [2], LiTS [1], MSD [23], KiTS [8,9], autoPET [6,7], TotalSegmentator [28], and AbdomenCT-1K [19]. The training set includes 4000 abdomen CT scans where 2200 CT scans with partial labels and 1800 CT scans without labels. The validation and testing sets include 100 and 400 CT scans, respectively, which cover various abdominal cancer types, such as liver cancer, kidney cancer, pancreas cancer, colon cancer, gastric cancer, and so on. The organ annotation process used ITK-SNAP [29], nnU-Net [11], and MedSAM [16].

The evaluation metrics encompass two accuracy measures-Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD)-alongside two efficiency measures-running time and area under the GPU memory-time curve. These metrics collectively contribute to the ranking computation. Furthermore, the running time and GPU memory consumption are considered within tolerances of 15 s and 4 GB, respectively.

3.2 Implementation Details

Environment Settings. The development environments and requirements are presented in Table 3.

Table 3. Development environments and requirements.

System	Ubuntu 20.04.5 LTS
CPU	Intel(R) Xeon(R) Gold 6354 CPU @ 3.00 GHz
RAM	16 ×4 GB; 1600 MT/s
GPU (number and type)	1 × NVIDIA A100 40 G
CUDA version	11.7
Programming language	Python 3.10.11
Deep learning framework	Pytorch 2.0.0, torchvision 0.2.2
Specific dependencies	nnU-Net 2.0
Code	https://github.com/orangeqqq/FLARE23

Training Protocols. The training protocols of the small nnU-Net are listed in Table 4. For the unlabeled images, we select 100 cases with the pseudo label to train the network. For partial labels, we use the partial cross-entropy and dice loss in the training stage. the pseudo labels generated by the FLARE22 winning algorithm [10] and the best-accuracy-algorithm [27]. We employ the same

Table 4. Training protocols.

Network initialization	"He" normal initialization
Batch size	4
Patch size	128 × 128 × 128
Total epochs	1000
Optimizer	SGD with nesterov momentum (μ =0.99)
Initial learning rate (lr)	0.01
Lr decay schedule	Poly learning rate policy: $(1 - epoch/1000)^{0.9}$
Training time	10 h
Loss function	Cross entropy loss and dice loss
Number of model parameters	5.22 M[a]
Number of flops	121 G[b]
CO_2eq	11.2 Kg[c]

[a] https://github.com/sksq96/pytorch-summary
[b] https://github.com/facebookresearch/fvcore
[c] https://github.com/lfwa/carbontracker/

data augmentation as the default setting of nnU-Net, which includes additive brightness, gamma, rotation, scaling, and elastic deformation on the fly during training. During inference, the model does not perform test time augmentation (TTA) of flipping. The patch sampling strategy is foreground over-sampling. Finally, we choose the model that obtains the fast and best accuracy on the online validation.

4 Results and Discussion

Table 5. Quantitative evaluation results in terms of DSC(%) and NSD(%).

Target	Public Validation		Online Validation		Testing	
	DSC	NSD	DSC	NSD	DSC	NSD
Liver	95.54 ± 2.53	96.86 ± 5.34	95.62	97.12	93.65	95.47
Right Kidney	87.89 ± 19.54	88.35 ± 20.41	89.35	90.01	91.63	91.62
Spleen	93.06 ± 3.77	93.55 ± 8.12	93.18	93.86	92.49	93.09
Pancreas	82.05 ± 5.93	95.41 ± 4.86	80.72	94.5	82.06	95.09
Aorta	93.05 ± 2.06	97.64 ± 3.19	93.35	97.98	93.04	98.29
Inferior vena cava	88.05 ± 5.56	90.98 ± 6.52	88.06	90.7	88.69	92.17
Right adrenal gland	74.67 ± 12.86	91.33 ± 13.74	75.24	92.12	72.23	90.71
Left adrenal gland	71.41 ± 13.29	88.43 ± 14.0	72.83	89.22	71.06	88.45
Gallbladder	82.06 ± 19.92	81.27 ± 21.06	82.52	81.86	74.11	74.13
Esophagus	78.46 ± 14.01	91.15 ± 14.41	79.12	92.15	81.85	94.98
Stomach	90.23 ± 6.08	95.25 ± 6.71	90.6	95.07	89.48	94.36
Duodenum	78.06 ± 8.28	93.96 ± 5.57	78.25	93.53	78.86	94.19
Left kidney	86.96 ± 16.61	87.77 ± 17.72	87.96	88.78	91.23	91.43
Tumor	18.21 ± 23.28	10.27 ± 15.24	15.14	8.72	17.62	8.37
Average	79.98 ± 10.98	85.87 ± 11.21	80.14	86.12	79.86	85.88

4.1 Quantitative Results on Validation Set

In Table 5, we report the DSC and NSD of the final docker commit results. The average of the 50 public validation and the 100 online validation are the same, both achieving a DSC of about 0.80 and an NSD of 0.86. In general, large organs like the liver, spleen, kidney, and stomach have high accuracy. However, accurate identification of small and complex objects, such as tumors, adrenal glands, and the duodenum, presents significant challenges. It requires more attention, especially when dealing with extremely small and indistinct boundaries.

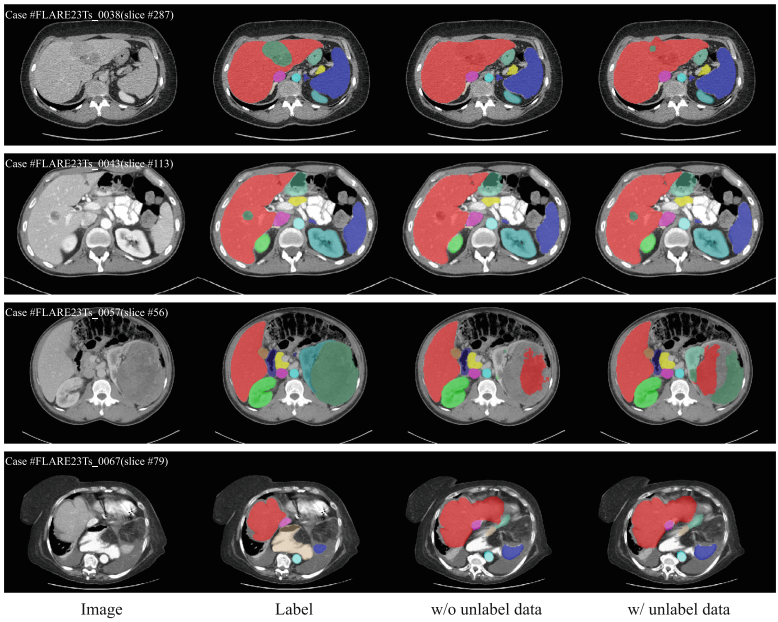

Fig. 2. Qualitative results on two easy cases (Case #FLARE23Ts_0038 with DSC of 0.89 and Case #FLARE23Ts_0043 with DSC of 0.84) and two hard cases (Case #FLARE23Ts_0057 with DSC of 0.66 and Case #FLARE23Ts_0067 with DSC of 0.74).

We report the online validation results of the model without unlabelled data, normal inference processes, and cascade networks in Table 7. The model using unlabelled data resulted in an increase of the DSC from 0.7925 to 0.8013. Specifically, in tumor regions, it increased the DSC by 0.045. Additionally, normal inference alone increased the DSC by approximately 0.04. However, the cascade network, P-Cascade and N-Cascade, which added a network training in a lower resolution setup with twice the spacing of the original, did not achieve higher DSC and NSD results. P-Cascade is the results of partially compound loss and N-Cascade is the results of normal compound loss. Comparing the two, we find

that the model trained by partially labeled loss has better results for tumor segmentation, with an improvement in DSC value of 0.05.

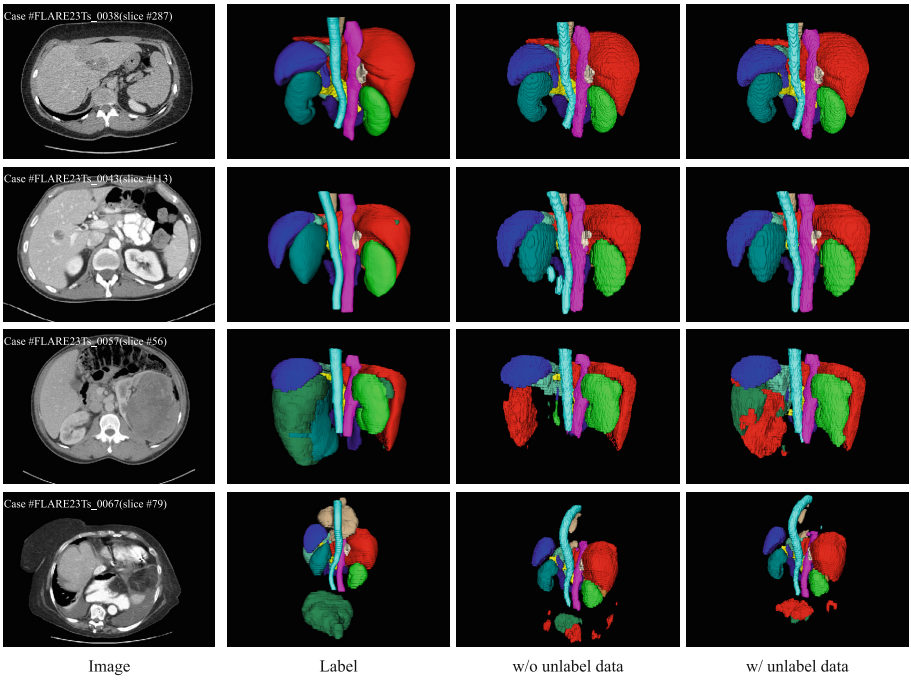

Fig. 3. 3D visualization on two easy cases (Case #FLARE23Ts_0038 with DSC of 0.89 and Case #FLARE23Ts_0043 with DSC of 0.84) and two hard cases (Case #FLARE23Ts_0057 with DSC of 0.66 and Case #FLARE23Ts_0067 with DSC of 0.74).

4.2 Qualitative Results on Validation Set

Figure 2 presents easy and difficult validation set examples for segmentation, along with a 3D visualization in Fig. 3. Promising results were observed for Case #FLARE23Ts_0038 and Case #FLARE23Ts_0043, but the segmentation of Case #FLARE23Ts_0057 and Case #FLARE23Ts_0067 was poor due to a large tumor that caused the network to make classification errors.

4.3 Segmentation Efficiency Results on Validation Set

In Table 6, we observe a set of cases that increase in size from (512, 512, 55) to (512, 512, 554). The efficiency evaluation results are reported from official tests. It is seen that the average max GPU is 2531MB, and run time increase twice for the biggest case #0029 than the smallest case #0001. This demonstrates the effectiveness of our inference strategy.

Table 6. Quantitative evaluation of segmentation efficiency in terms of the running time and GPU memory consumption. Total GPU denotes the area under GPU Memory-Time curve. Evaluation GPU platform: NVIDIA QUADRO RTX5000 (16 G).

Case ID	Image Size	Running Time (s)	Max GPU (MB)	Total GPU (MB)
0001	(512, 512, 55)	19.61	2426	10028
0051	(512, 512, 100)	17.83	2590	12296
0017	(512, 512, 150)	30.86	2634	15949
0019	(512, 512, 215)	22.72	2486	12401
0099	(512, 512, 334)	27.94	2586	15394
0063	(512, 512, 448)	33.50	2630	17508
0048	(512, 512, 499)	35.22	2614	18610
0029	(512, 512, 554)	42.53	2744	22299

4.4 Results on Final Testing Set

In Table 5, we report the DSC and NSD of the final testing set. The average values are comparable to those of the 50 public validations and the 100 online validations, with both achieving a DSC of about 0.80 and a NSD of about 0.86. In general, the low accuracy of segmenting small and complex shaped objects such as tumors, adrenal glands and duodenums Their accurate segmentation still faces great challenges and needs more attention, especially when dealing with extremely small and unclear boundaries.

4.5 Limitation and Future Work

There are many ways to improve the network inference process, such as a more efficient sliding window. The challenge provided 4000 CT cases, but we only utilized 300 cases and did not adequately utilize the data. For the challenging task of tumor segmentation, pseudo-labeling is a simple and effective way to improve model performance, and we will continue to explore methods that utilize both pseudo-labeling and partial labeling learning in the future.

Table 7. Ablation studies of online validation quantitative evaluation results in terms of DSC(%) and NSD(%). P-Cascade is the results of partially compound loss and N-Cascade is the results of normal compound loss.

Target	w/o unlabeled data		Normal inference		N-Cascade		P-Cascade	
	DSC	NSD	DSC	NSD	DSC	NSD	DSC	NSD
Liver	95.77	97.09	97.34	97.46	95.63	97.63	95.9	97.5
Right Kidney	89.93	90.49	92.18	91.46	90.27	91.27	89.9	91.28
Spleen	93.57	94.46	97	97.58	91.34	92.01	92.68	93.44
Pancreas	79.66	93.5	84.22	94.82	79.74	93.78	79.79	93.6
Aorta	92.29	96.86	96.57	99.03	92.59	97.42	93.23	97.79
Inferior vena cava	87.24	89.84	91.06	91.43	86.38	88.25	87.06	89.12
Right adrenal gland	74.24	91.78	85.51	95.48	72.75	90.19	73.35	90.59
Left adrenal gland	71.19	87.59	83.27	93.27	72.47	89.09	72.33	88.76
Gallbladder	80.34	79.38	86.09	86.55	77.9	77.05	80.54	79.83
Esophagus	78.13	90.88	83.09	93.4	78.57	91.89	79.05	92.26
Stomach	90.52	94.52	93.12	95.51	89.95	94.58	90.37	94.91
Duodenum	77.31	93.19	81.45	93.43	78.42	94.19	78.25	93.97
Left kidney	88.69	88.97	91.06	90.65	88.23	89.43	87.67	86.93
Tumor	10.64	5.92	15.17	8.42	10.25	6.99	15.88	10.43
Average	79.25	85.32	84.08	87.75	78.89	85.27	79.71	85.74

5 Conclusion

In this paper, we present a framework that combines partial labeling learning and pseudo-labeling, which is effective and flexible for a variety of situations. In addition, we use a small nnU-Net and improve the inference process, effectively reducing its required computational resources and inference time. Because the amount of data used in training is small, performance on the full data will be explored in the future. The approach in this paper will be a good baseline result for exploring partial labeling learning and pseudo-labeling.

Acknowledgements. The authors of this paper declare that the segmentation method they implemented for participation in the FLARE 2023 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. We thank all the data owners for making the CT scans publicly available and CodaLab [20] for hosting the challenge platform.

References

1. Bilic, P., et al.: The liver tumor segmentation benchmark (lits). Med. Image Anal. **84**, 102680 (2023)
2. Clark, K., et al.: The cancer imaging archive (TCIA): maintaining and operating a public information repository. J. Digit. Imaging **26**(6), 1045–1057 (2013)
3. Dmitriev, K., Kaufman, A.E.: Learning multi-class segmentations from single-class datasets. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

4. Fang, X., Yan, P.: Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. IEEE Trans. Med. Imaging **39**(11), 3619–3629 (2020)
5. Fidon, L., et al.: Label-set loss functions for partial supervision: application to fetal brain 3d MRI parcellation. In: Medical Image Computing and Computer Assisted Intervention (2021)
6. Gatidis, S., et al.: The autopet challenge: towards fully automated lesion segmentation in oncologic pet/ct imaging. preprint at Research Square (Nature Portfolio) (2023). https://doi.org/10.21203/rs.3.rs-2572595/v1
7. Gatidis, S., et al.: A whole-body FDG-pet/CT dataset with manually annotated tumor lesions. Sci. Data **9**(1), 601 (2022)
8. Heller, N., et al.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: results of the kits19 challenge. Med. Image Anal. **67**, 101821 (2021)
9. Heller, N., 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. Proc. Am. Soc. Clin. Oncol. **38**(6), 626–626 (2020)
10. Huang, Z., et al.: Revisiting nnU-net for iterative pseudo labeling and efficient sliding window inference. In: Ma, J., Wang, B. (eds.) Fast and Low-Resource Semi-supervised Abdominal Organ Segmentation, FLARE 2022, LNCS, vol. 13816, pp. 178–189. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-23911-3_16
11. 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. Nat. Methods **18**(2), 203–211 (2021)
12. Li, S., Wang, H., Meng, Y., Zhang, C., Song, Z.: Multi-organ segmentation: a progressive exploration of learning paradigms under scarce annotation. arXiv preprint arXiv:2302.03296 (2023)
13. Liu, J., et al.: Clip-driven universal model for organ segmentation and tumor detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 21152–21164 (2023)
14. Luo, X., et al.: Word: a large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from CT image. Med. Image Anal. **82**, 102642 (2022)
15. Ma, J., et al.: Loss odyssey in medical image segmentation. Med. Image Anal. **71**, 102035 (2021)
16. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nat. Commun. **15**(1), 654 (2024)
17. Ma, J., et al.: Fast and low-GPU-memory abdomen CT organ segmentation: the flare challenge. Med. Image Anal. **82**, 102616 (2022)
18. Ma, J., et al.: Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. arXiv preprint arXiv:2308.05862 (2023)
19. Ma, J., et al.: Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Trans. Pattern Anal. Mach. Intell. **44**(10), 6695–6714 (2022)
20. Pavao, A., et al.: Codalab competitions: an open source platform to organize scientific challenges. J. Mach. Learn. Res. **24**(198), 1–6 (2023)
21. Roulet, N., Slezak, D.F., Ferrante, E.: Joint learning of brain lesion and anatomy segmentation from heterogeneous datasets. In: Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning (2019)
22. Shi, G., Xiao, L., Chen, Y., Zhou, S.K.: Marginal loss and exclusion loss for partially supervised multi-organ segmentation. Med. Image Anal. **70**, 101979 (2021)

23. Simpson, A.L., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019)
24. Tang, Y., et al.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
25. Ulrich, C., Isensee, F., Wald, T., Zenk, M., Baumgartner, M., Maier-Hein, K.H.: Multitalent: a multi-dataset approach to medical image segmentation. arXiv preprint arXiv:2303.14444 (2023)
26. Wang, C., Cui, Z., Yang, J., Han, M., Carneiro, G., Shen, D.: Bowelnet: joint semantic-geometric ensemble learning for bowel segmentation from both partially and fully labeled CT images. IEEE Trans. Med. Imaging **42**(4), 1225–1236 (2023)
27. Wang, E., Zhao, Y., Wu, Y.: Cascade dual-decoders network for abdominal organs segmentation. In: Ma, J., Wang, B. (eds.) Fast and Low-Resource Semi-supervised Abdominal Organ Segmentation, FLARE 2022, LNCS, vol. 13816, pp. 202–213. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-23911-3_18
28. Wasserthal, J., et al.: Totalsegmentator: robust segmentation of 104 anatomic structures in CT images. Radiol. Artif. Intell. **5**(5), e230024 (2023)
29. Yushkevich, P.A., Gao, Y., Gerig, G.: Itk-snap: an interactive tool for semi-automatic segmentation of multi-modality biomedical images. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 3342–3345 (2016)
30. Zhang, J., Xie, Y., Xia, Y., Shen, C.: Dodnet: learning to segment multi-organ and tumors from multiple partially labeled datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021
31. Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J.: Models genesis. Med. Image Anal. **67**, 101840 (2021)