3D Swin Transformer for Partial Medical Auto Segmentation

Aneesh Rangnekar[0000−0002−0079−9495], Jue Jiang, Harini Veeraraghavan

Memorial Sloan Kettering Cancer Center
rangnea@mskcc.org

Abstract. Transformers are the highest accuracy segmentation frameworks in computer vision for natural imagery from the past few years. In contrast, medical imaging approaches, except a select few (for example, SwinUNETR and SMIT), are still dominated by the nnU-Net architecture family. In this paper, we investigate the application of a hierarchical vision transformer to the FLARE-23 challenge.
Specifically, we benchmark our results using a relatively lightweight architecture, Swin-X Seg. We use multi-model self-training, wherein we use nnU-Net for predicting pseudo labels on partially labeled cases and then optimize the transformer architecture for memory requirements. Our network achieved the average DSC scores of 83.13 % and 35.19 % on the open validation set (50 cases) for organs and tumors, respectively, while staying under a max GPU memory utilization of 4GB at evaluation runtime. Our results show that there is potential for the transformer architecture to perform at par or better than conventional convolutional approaches, and we hope our findings encourage more research in the area.

Keywords: Auto Segmentation · Self-training · Swin Transformer.

1 Introduction

Accurate, fast, and automated volumetric segmentation of organs and tumors is essential for radiotherapy treatment planning. It often constitutes one of the time-consuming parts of radiation treatment planning workflows [37]. Abdominal organs are particularly time-consuming to segment owing to the presence of a large number of organs as well as due to the random and large variation in the appearance and shape of gastrointestinal organs and limited soft-tissue contrast on clinically used computed tomography (CT) images. Hence, deep learning methods to generate segmentation are under active development [20,2].

Deep learning methods have shown the capability to generate multi-organ segmentation for abdomen [16,18,34,1] and other disease sites. The availability of well-curated public challenge datasets [20,2] has enabled the evaluation of various methods using the same reference benchmark with well-defined metrics. However, a fundamental prerequisite of well-curated pixel-wise annotations or volumetric segmentations of the various organs for training these networks must

be more expensive and time-consuming to generate on large datasets. One recent promising approach to alleviate the need for large, curated datasets is the self-supervised pretraining followed by a fine-tuning approach that has demonstrated success in medical image analysis, mainly when using transformer-based architectures[34,18]. Swin UNETR [34] and SMIT [18] have shown that using self-supervised learning (SSL) improves the performance of transformer-based networks on semantic segmentation, as compared to training the networks from scratch. Our approach builds on these methods and utilizes a transformer architecture [21] for segmentation with a pretraining step (self-supervised learning) using labeled and unlabeled examples followed by fine-tuning.

We also follow the FLARE-23 rules, whereby, unlike prior works[34,18], which used a large number of CT scans from various disease sites for pretraining, we used only the 4,000 example scans provided as part of the training set for self-supervised pretraining. Furthermore, keeping with the requirements for using a relatively small architecture with limited memory requirements, we also constructed a lightweight transformer architecture.

Our learning framework uses multi-model self-training [41,42,32], where the teacher is an fine-tuned nnU-Net [15] that generates pseudo labels for the various categories. The student network uses a Swin transformer backbone [21] segmentation network (here on referred to as Swin-X Seg) that accepts a combination of FLARE-23 and pseudo labeled examples for fine-tuning (Fig. 1). Our initial studies show that naively using the partially labeled dataset, with a transformer backbone to obtain pseudo labels, results in poor performance across multiple categories [36,5,40]. . Hence, we resort to this combination of semi-supervised learning, wherein the teacher is an nnU-Net and the student is Swin-X Seg.

Our approach allows us to fully utilize the partially-labeled training dataset to its fullest extent, while leveraging fundamental augmentation techniques shown to be effective in natural image analysis. This mitigates the need for requiring complex approaches like the CutMix [43] or ClassMix [29], wherein extensive registration would be required before mixing two 3D scans so that the networks do not lose understanding of organ placements, especially with architectures that rely heavily on positional information.

Our key contributions are (a) a lightweight 3D vision transformer applied to multi-organ and tumor segmentation, (b) the SSL approach extending prior works by learning the downstream task using partial labels, and the application of this approach on an open-source FLARE-23 dataset.

2 Method

2.1 Overview

We studied the performance of hierarchical vision transformer-based U-Net architecture on the FLARE-23 challenge. Vision transformers require large amounts of data [36,5,40,19] to achieve high generalization performance. Hence, FLARE-23, which consists of 4,000 training images, provides a nice test bed for evaluating

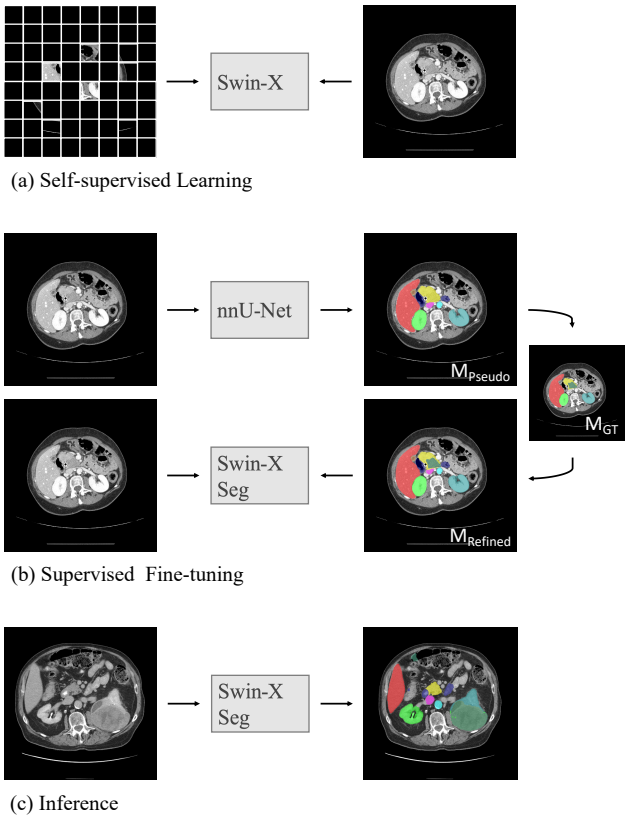

Fig. 1. Our three-stage pipeline: (a) self-supervised training of the backbone network [17], (b) uses a combination of pseudo labels (M_{Pseudo}) [15] and FLARE-23 provided annotations (M_{GT}) to obtain refined labels ($M_{Refined}$) for learning segmentation, and (c) inference on a new unseen volumetric scan.

vision transformer architectures. However, 1800 CTs in FLARE-23 are unlabeled with the remaining 2200 CTs provided with partial labels, wherein some but not all the 14 different organs and tumors were segmented, which makes supervised training challenging. Therefore, we used a two-step training approach consisting of: (i) self-supervised pretraining performed on the entire dataset of 4,000 CTs without using any segmentations for supervised training, and (ii) supervised fine-tuning that combined fully labeled CTs together with CTs with pseudo labels created using a different model. We discuss each part of our approach in detail, and the specificities involved in our final implementation.

2.2 Preprocessing

We used the following preprocessing steps in all our experiments:

– Reorient the scans to the right-anterior-superior (RAS) view.

- Clip the intensities based on the Hounsfield units to [-250, 250].
- We resize all scans to x, y, z volumetric spacings of $1.0, 1.0, 1.0$ during training and inference.
- In addition, we randomly sample 4 scans of $96 \times 96 \times 96$ size from each scan as training examples, representing 2 positive and 2 negative samples for the network at every instance.

2.3 Proposed Method

Choice of Transformer:

Hierarchical Vision Transformers [21,8] are pyramid-shaped architectures that rely on gradual down-sampling, similar to convolutional neural networks, while maintaining a global look-out with their multi-scale designs. We use the Swin-Transformer backbone for our approach as it has been widely adopted for 3D medical auto segmentation [34,18] and shown to be more accurate than the vanilla vision transformer[7].

Swin UNETR [34] and SMIT [18] have over 60 million (M) parameters. Whereas Swin UNETR processes data at $96 \times 96 \times 96$, SMIT processes data at $128 \times 128 \times 128$ resolution. Both methods use sliding windows for generating final inference. The FLARE-23 constraints require memory efficient inference. A straightforward memory efficient approach to reduce the total number of flops used for inference would be to utilize CT scans reduced to $96 \times 96 \times 96$ pixels, at the risk of decreasing the image resolution, which can impact accuracy for smaller organs. Hence, we reduced the number of parameters used in the network by decreasing the total number of blocks per depth to the final $2-2-2-2$ configuration as well as reduced the total number of channels through the UNETR architecture using 1×1 convolutions. This reduced the network size from 60M parameters to 31M parameters, a relatively lightweight architecture compared to current state-of-the-art methods. This is also crucial towards keeping the GPU requirements under 4GB as stipulated under FLARE-23 rules.

Self-supervised Learning: The SSL approach made use of the self-distillation based pretext tasks used in the SMIT [18], including namely Masked Image Modeling (MIM), Masked Patch self-Distillation (MPD) and Image Token self-Distillation (ITD). SMIT performs self distillation by concurrently maintaining an online teacher model (NET_T) with the same network architecture as the student model (NET_S) [35]. The loss functions used to optimize the network are briefly discussed here and we refer interested details to the original paper[18] for more details.

Suppose $\{x_1, x_2\}$ are two augmented views of a 3D image x. N image patches are extracted from the images to create a sequence of image tokens [7]. The image tokens are then corrupted by randomly masking image tokens based on a binary vector, with a probability p, and then replacing with mask token [3]. The second augmented view v is also corrupted but using a different mask vector instance. In this order, the three losses deal with the views in the following manner:

- **Masked Image Prediction (MIP)** $\rightarrow x_1$, NET_S, involves dense pixel regression of image intensities within masked patches using the context of unmasked patches [12].
- **Masked patch token self-distillation (MPD):** $\rightarrow x_1$, NET_S, NET_T, trains the student network to predicts the tokens of the teacher network (distillation).
- **Global image token self-distillation (ITD):** $\rightarrow x_1, x_2$, NET_S, NET_T, learns to match the global image embedding of the view-scan seen by the student network to the view-scan seen by the teacher network.

SSL training is performed by optimizing the network using all three aforementioned losses. FLARE-23 rules dictate that no external data be used. Hence, following the rules, SSL used the same 4,000 CTs provided as part of the training set. No segmentations provided with the data was used for network optimization in this step.

Supervised Fine-tuning:

In order to fully utilize all available training data to improve accuracy, we used the best performing nnU-Net model, the winner from FLARE22[15] to provide pseudo labels for the partially labeled and unlabeled datasets the FLARE 23 training sets. We only use 735 examples from the 2200 images that contain a labeled instance of tumor, with the combination of FLARE-23 and nnU-Net pseudo labels (Fig. 1). We trained our network sing a combination of Dice loss and cross-entropy loss following previous approaches [24,16,34,18].

2.4 Post-processing

No data specific post processing was used following pixel-level classifications generated by the segmentation methods. Sliding window inference with 50% overlap was used for generating segmentations for the whole 3D image volumes.

3 Experiments

3.1 Dataset and evaluation measures

The FLARE-23 challenge is an extension of the FLARE 2021-2022 [26][27], aiming to promote the development of foundation models in abdominal disease analysis. The segmentation targets cover 13 organs and various abdominal lesions around the organs. The dataset comprises scans from more than 30 medical centers, including TCIA [6], LiTS [4], MSD [33], KiTS [13,14], autoPET [10,9], TotalSegmentator [39], and AbdomenCT-1K [28], with appropriate licensing. The training set includes 4,000 abdomen CT scans, 2,200 CT scans with partial segmentation labels for some of them, and 1,800 CT scans without any segmentation labels. The validation and testing sets include 100 and 400 CT scans, respectively, covering various abdominal cancer types, such as liver, kidney, pancreas, colon, and gastric, to name a few. The organ annotation process used ITK-SNAP [44], nnU-Net [16], and MedSAM [25].

Table 1. Development environments and requirements.

System	Ubuntu 18.04.5 LTS
CPU	AMD EPYC 7543P 32-Core Processor @ 2.8 Ghz
RAM	128 GB
GPU (number and type)	NVIDIA A100 80 GB × 4
CUDA version	11.8
Programming language	Python 3.8
Deep learning framework	Pytorch 1.13 ± CUDA 11.7 [30]
Specific dependencies	MONAI, SimpleITK, Nibabel
Code	https://github.com/The-Veeraraghavan-Lab/FLARE23

Table 2. Training protocols.

Network initialization	SSL-FLARE-23 [18]
Batch size	4
Patch size	96 × 96 × 96
Total epochs	100
Optimizer	AdamW [23]
Initial learning rate (lr)	$2e$-4
Lr decay schedule	Linear Warmup with Cosine Annealing [22,11]
Training time	33 hours
Loss function	Cross-Entropy Loss /w Dice Loss

The evaluation metrics encompass two accuracy measures—Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD)—alongside two efficiency measures—running time and instantaneous GPU maximum memory consumption.

3.2 Implementation details

Environment settings The development environments and requirements are presented in Table 1. We provide all the requirements in our released codebase on GitHub.

Training protocols The model training protocols are shown in in Table 2. An image patch size of 96 × 96 × 96 with random 3D flips performed on the data to provide augmented samples was used for network training.

Table 3. Quantitative evaluation results. Segmentation accuracy results (DSC and NSD with mean and standard deviation) are reported on the publicly provided 50 validation cases made available by the FLARE-23 organizers.

Target	Public Validation	
	DSC(%)	NSD(%)
Liver	96.08 ± 4.230	93.58 ± 10.66
Right Kidney	87.00 ± 20.81	83.37 ± 21.81
Spleen	93.24 ± 9.730	90.92 ± 14.23
Pancreas	80.47 ± 7.860	89.99 ± 7.020
Aorta	90.55 ± 14.80	91.61 ± 16.30
Inferior vena cava	87.88 ± 6.800	86.97 ± 9.300
Right adrenal gland	77.35 ± 17.46	87.78 ± 19.00
Left adrenal gland	72.44 ± 15.83	82.03 ± 16.59
Gallbladder	75.61 ± 28.21	71.61 ± 30.06
Esophagus	74.81 ± 16.56	84.85 ± 15.99
Stomach	89.17 ± 9.110	87.60 ± 11.85
Duodenum	70.78 ± 10.77	84.21 ± 9.240
Left kidney	85.65 ± 21.81	82.33 ± 23.22
Tumor	35.19 ± 30.17	22.99 ± 22.10
Average (Organ)	83.13 ± 8.440	85.55 ± 12.58
Average	79.70 ± 11.43	81.08 ± 14.93

4 Results and discussion

4.1 Quantitative results on validation set

Table 3 shows our Swin-X Seg's performance on the 50 validation cases provided by the FLARE-23 organizers. The network was slightly less accurate ($<$ 80% DSC) for organs such as the adrenal glands, gallbladder, esophagus, duodenum, as well as for tumors compared to larger organs like the liver, spleen, left and right kidneys, and the stomach. The tumor segmentation accuracy was low because of the larger variability in the types of tumors analyzed and the relatively few examples with complete labels. Overall, the network accuracy was lower for smaller organs like the adrenal glands and gallbladder when compared to larger organs like the liver. Poor accuracy for organs also resulted when they were adjacent to the tumors.

Table 4 shows that inference requirements of under 4GB GPU memory consumption were satisfied for all cases. However, all except two cases (0001, 0019) did not satisfy the running time requirement under 60 secs owing to sliding window-based inference, with 50% overlap. A natural option is to use sliding window inference without any overlap (0%). However, this results in a poor overall score (77% DSC average on organ, 27% DSC on tumor); hence, we did

Table 4. Quantitative evaluation of segmentation efficiency of the reported cases using running time and maximum GPU memory consumption (< 4096 MB). Evaluation GPU platform: A100 (80GB).

Case ID	Image Size	Running Time (s)	Max GPU (MB)
0001	(512, 512, 55)	28.01	3464
0051	(512, 512, 100)	65.86	3850
0017	(512, 512, 150)	73.94	3896
0019	(512, 512, 215)	48.00	3616
0099	(512, 512, 334)	69.28	3756
0063	(512, 512, 448)	84.76	3776
0048	(512, 512, 499)	74.73	3748
0029	(512, 512, 554)	102.5	4032

not pursue it. In addition, we optimized for test-time efficiency by performing foreground thresholding to use only the body regions for analysis by ignoring the surrounding air for inference. Our analysis showed that in cases with larger field of view, wherein the body occupied higher volume the inference time utilization increased (e.g. 0017 > 0019, 0063 > 0048).

4.2 Qualitative results on validation set

Figures 2 and 3 show the segmentations generated by our network on representative examples taken from the validation set of FLARE-23. As shown in Fig.2, whereas the model tends to consistently segment the normal tissues with high accuracy, misclassifications occur within tumor regions, tumor voxels classified as the kidney, despite achieving a relatively high DSC accuracy for the tumors. The higher DSC accuracy for tumors is not surprising given the larger tumor volumes. On the other hand, as shown in Fig. 3 for really large tumors such as #0057 and #0095, the algorithm generated highly inaccurate segmentation, misclassifying the tumors occurring on the left side of anatomy as liver. #0027 shows an example where the kidney tumor was correctly segmented together with the kidney adjacent to the tumor, although the esophagus occurring distally to the pancreatic head was misclassified as pancreas. Similarly, in #0089, the pancreas is oversegmented by the model, whereas the kidney tumor encased within the kidney is undersegmented, highlighting the challenges, particularly when the tumor and the healthy tissues are adjacent to each other.

4.3 Segmentation efficiency results on validation set

We optimized for segmentation inference efficiency by extracting the foreground or the body as a preprocessing step using standard image thresholding. No additional optimization was performed in terms of training or testing. Even this simple approach showed that it is possible to improve inference efficiency as seen in Table 4.

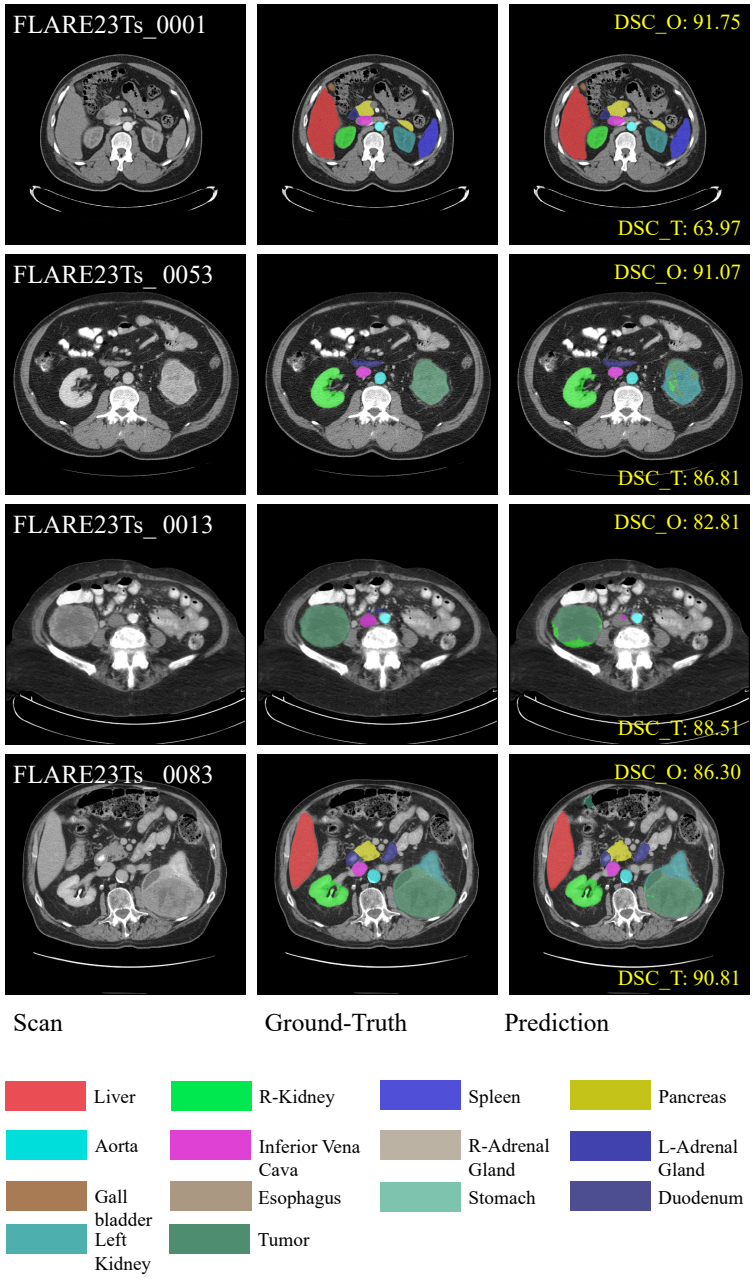

Fig. 2. Example scans showing relatively good performance in terms of misclassifications by the trained Swin-X Seg model. DSC_T refers to tumor DSC and DSC_O refers to average multi-organs DSC.

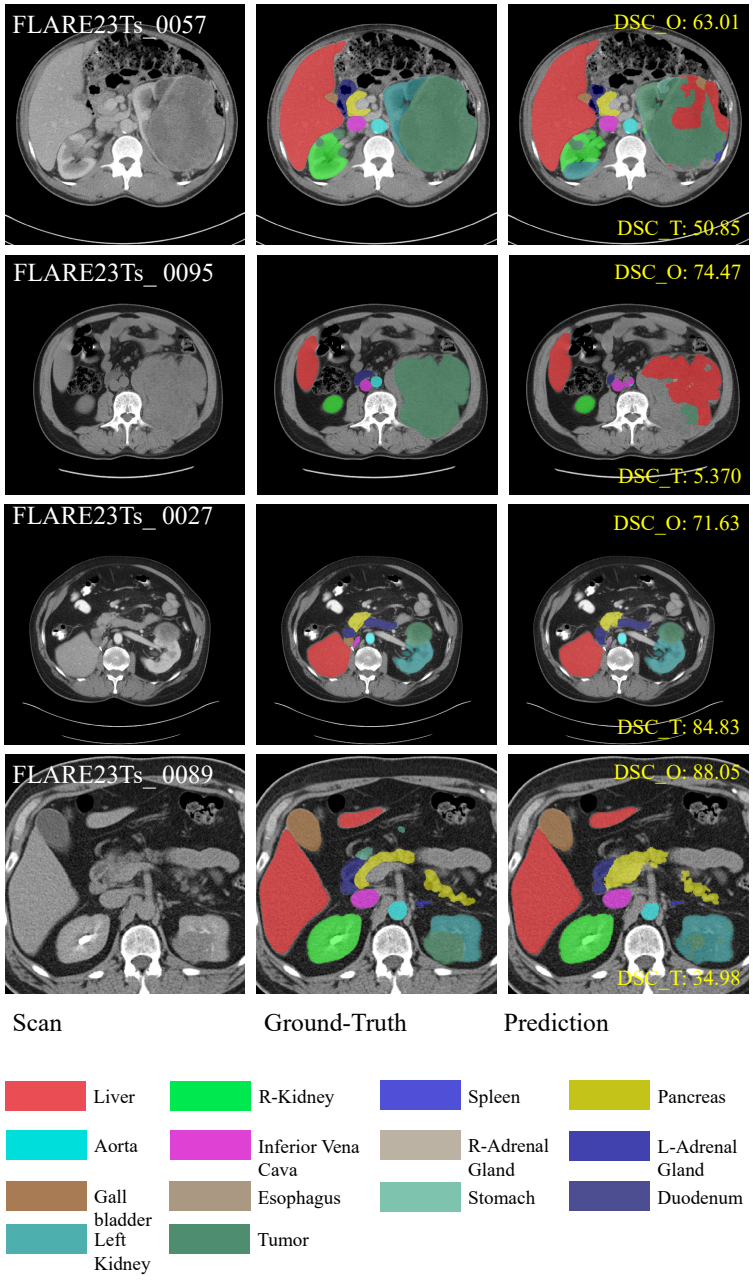

Fig. 3. Example scans showing relatively poor performance in terms of misclassifications by the trained Swin-X Seg network. DSC_T refers to tumor DSC and DSC_O refers to average multi-organs DSC.

4.4 Results on final testing set

This is a placeholder. We will send you the testing results during MICCAI (2023.10.8). (This is to be left as is.)

4.5 Limitation and future work

Our goal was to evaluate the capability of transformer-based approach for multi-organ and tumor segmentation. We used a relatively lightweight (31M) in order to satisfy the memory requirements of the competition as well as to study to what extent such methods are successful in comparison to convolutional-based approaches such as the nnU-Net used in the previous iteration of the competition [15,38]. Our approach to use nnU-Net generated pseudo labels was motivated by prior results using Semiformer [40], which showed poor accuracy with vision transformer with small labeled training samples can be improved when combined with pseudo labels produced by convolutional neural networks (CNN). However, VITs have generally shown to be more accurate than CNN models. Hence, one approach is to use VIT instead of a CNN for providing pseudo labels. its important to note that the approach combining pseudo labels with CNN and larger VIT models becomes impractical due to increasing memory needs. Another limitation of our approach is the poor segmentations we observed on the tumor and tissue interface, which we plan to address in the future.

5 Conclusion

We presented our approach, multi-model self-training, that used nnU-Net to generate pseudo labels and then Swin transformer to establish a foundation for research into auto segmentation with pseudo labels. In addition, we also identify limitations and discuss research approaches to mitigate them, including knowledge distillation and semi-supervised learning. We believe that our framework serves as a good foundation for further research into efficient network designs and methodology for accurate medical image segmentation.

Acknowledgements The authors of this paper declare that the segmentation method they implemented for participation in the FLARE-23 challenge has not used any pre-trained models and 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 [31] for hosting the challenge platform. This research was partly funded through grant from NCI R01CA258821-01A1 and the Memorial Sloan Kettering (MSK) Cancer Center Support Grant/Core Grant NCI P30 CA008748.

References

1. Amjad, A., Xu, J., Thill, D., Lawton, C., Hall, W., Awan J., M., Shukla, M., Erickson, B.A., Li, X.A.: General and custom deep learning autosegmentation models for organs in head and neck, abdomen, and male pelvis. Med Phys **49**(3), 1686–1700 (2022) 1
2. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022) 1
3. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021) 4
4. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., Lohöfer, F., Holch, J.W., Sommer, W., Hofmann, F., Hostettler, A., Lev-Cohain, N., Drozdzal, M., Amitai, M.M., Vivanti, R., Sosna, J., Ezhov, I., Sekuboyina, A., Navarro, F., Kofler, F., Paetzold, J.C., Shit, S., Hu, X., Lipková, J., Rempfler, M., Piraud, M., Kirschke, J., Wiestler, B., Zhang, Z., Hülsemeyer, C., Beetz, M., Ettlinger, F., Antonelli, M., Bae, W., Bellver, M., Bi, L., Chen, H., Chlebus, G., Dam, E.B., Dou, Q., Fu, C.W., Georgescu, B., i Nieto, X.G., Gruen, F., Han, X., Heng, P.A., Hesser, J., Moltz, J.H., Igel, C., Isensee, F., Jäger, P., Jia, F., Kaluva, K.C., Khened, M., Kim, I., Kim, J.H., Kim, S., Kohl, S., Konopczynski, T., Kori, A., Krishnamurthi, G., Li, F., Li, H., Li, J., Li, X., Lowengrub, J., Ma, J., Maier-Hein, K., Maninis, K.K., Meine, H., Merhof, D., Pai, A., Perslev, M., Petersen, J., Pont-Tuset, J., Qi, J., Qi, X., Rippel, O., Roth, K., Sarasua, I., Schenk, A., Shen, Z., Torres, J., Wachinger, C., Wang, C., Weninger, L., Wu, J., Xu, D., Yang, X., Yu, S.C.H., Yuan, Y., Yue, M., Zhang, L., Cardoso, J., Bakas, S., Braren, R., Heinemann, V., Pal, C., Tang, A., Kadoury, S., Soler, L., van Ginneken, B., Greenspan, H., Joskowicz, L., Menze, B.: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023) 5
5. Cao, Y.H., Yu, H., Wu, J.: Training vision transformers with only 2040 images. In: European Conference on Computer Vision. pp. 220–237. Springer (2022) 2
6. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., Prior, F.: The cancer imaging archive (tcia): maintaining and operating a public information repository. Journal of Digital Imaging **26**(6), 1045–1057 (2013) 5
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 4
8. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C.: Multiscale vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6824–6835 (2021) 4
9. Gatidis, S., Früh, M., Fabritius, M., Gu, S., Nikolaou, K., La Fougère, C., Ye, J., He, J., Peng, Y., Bi, L., 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/https://doi.org/10.21203/rs.3.rs-2572595/v1 5
10. Gatidis, S., Hepp, T., Früh, M., La Fougère, C., Nikolaou, K., Pfannenberg, C., Schölkopf, B., Küstner, T., Cyran, C., Rubin, D.: A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. Scientific Data **9**(1), 601 (2022) 5

11. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677 (2017) 6
12. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022) 5
13. Heller, N., Isensee, F., Maier-Hein, K.H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., Yao, G., Gao, Y., Zhang, Y., Wang, Y., Hou, F., Yang, J., Xiong, G., Tian, J., Zhong, C., Ma, J., Rickman, J., Dean, J., Stai, B., Tejpaul, R., Oestreich, M., Blake, P., Kaluzniak, H., Raza, S., Rosenberg, J., Moore, K., Walczak, E., Rengel, Z., Edgerton, Z., Vasdev, R., Peterson, M., McSweeney, S., Peterson, S., Kalapara, A., Sathianathen, N., Papanikolopoulos, N., Weight, C.: 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) 5
14. 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) 5
15. Huang, Z., Wang, H., Ye, J., Niu, J., Tu, C., Yang, Y., Du, S., Deng, Z., Gu, L., He, J.: Revisiting nnu-net for iterative pseudo labeling and efficient sliding window inference. In: MICCAI Challenge on Fast and Low-Resource Semi-supervised Abdominal Organ Segmentation. pp. 178–189. Springer (2022) 2, 3, 5, 11
16. 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) 1, 5
17. Jiang, J., Elguindi, S., Berry, S.L., Onochie, I., Cervino, L., Deasy, J.O., Veeraraghavan, H.: Nested block self-attention multiple resolution residual network for multiorgan segmentation from CT. Med Phys **49**(8), 5244–5257 (2022) 3
18. Jiang, J., Tyagi, N., Tringale, K., Crane, C., Veeraraghavan, H.: Self-supervised 3d anatomy segmentation using self-distilled masked image transformer (smit). In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 556–566. Springer (2022) 1, 2, 4, 5, 6
19. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023) 2
20. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. vol. 5, p. 12 (2015) 1
21. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021) 2, 4
22. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016) 6
23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 6

24. 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) 5
25. Ma, J., Wang, B.: Segment anything in medical images. arXiv preprint arXiv:2304.12306 (2023) 5
26. Ma, J., Zhang, Y., Gu, S., An, X., Wang, Z., Ge, C., Wang, C., Zhang, F., Wang, Y., Xu, Y., Gou, S., Thaler, F., Payer, C., Štern, D., Henderson, E.G., McSweeney, D.M., Green, A., Jackson, P., McIntosh, L., Nguyen, Q.C., Qayyum, A., Conze, P.H., Huang, Z., Zhou, Z., Fan, D.P., Xiong, H., Dong, G., Zhu, Q., He, J., Yang, X.: Fast and low-gpu-memory abdomen ct organ segmentation: The flare challenge. Medical Image Analysis **82**, 102616 (2022) 5
27. Ma, J., Zhang, Y., Gu, S., Ge, C., Ma, S., Young, A., Zhu, C., Meng, K., Yang, X., Huang, Z., Zhang, F., Liu, W., Pan, Y., Huang, S., Wang, J., Sun, M., Xu, W., Jia, D., Choi, J.W., Alves, N., de Wilde, B., Koehler, G., Wu, Y., Wiesenfarth, M., Zhu, Q., Dong, G., He, J., the FLARE Challenge Consortium, Wang, B.: Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. arXiv preprint arXiv:2308.05862 (2023) 5
28. 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) 5
29. Olsson, V., Tranheden, W., Pinto, J., Svensson, L.: Classmix: Segmentation-based data augmentation for semi-supervised learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1369–1378 (January 2021) 2
30. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf 6
31. Pavao, A., Guyon, I., Letournel, A.C., Tran, D.T., Baro, X., Escalante, H.J., Escalera, S., Thomas, T., Xu, Z.: Codalab competitions: An open source platform to organize scientific challenges. Journal of Machine Learning Research **24**(198), 1–6 (2023) 11
32. Rangnekar, A., Kanan, C., Hoffman, M.: Semantic segmentation with active semi-supervised representation learning. arXiv preprint arXiv:2210.08403 (2022) 2
33. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., Bilic, P., Christ, P.F., Do, R.K.G., Gollub, M., Golia-Pernicka, J., Heckers, S.H., Jarnagin, W.R., McHugo, M.K., Napel, S., Vorontsov, E., Maier-Hein, L., Cardoso, M.J.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019) 5
34. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medi-

cal image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20730–20740 (2022) 1, 2, 4, 5
35. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 1195–1204. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017) 4
36. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021) 2
37. Vandewinckele, L., Claessens, M., Dinkla, A., Brouwer, C., Crijns, W., Verellen, D., van Elmpt, W.: Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. Radiother Oncol **153**, 55–66 (2020) 1
38. Wang, E., Zhao, Y., Wu, Y.: Cascade dual-decoders network for abdominal organs segmentation. In: MICCAI Challenge on Fast and Low-Resource Semi-supervised Abdominal Organ Segmentation. pp. 202–213. Springer (2022) 11
39. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023) 5
40. Weng, Z., Yang, X., Li, A., Wu, Z., Jiang, Y.G.: Semi-supervised vision transformers. In: European Conference on Computer Vision. pp. 605–620. Springer (2022) 2, 11
41. 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) 2
42. Yang, L., Zhuo, W., Qi, L., Shi, Y., Gao, Y.: St++: Make self-training work better for semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 2
43. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6023–6032 (2019) 2
44. 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) 5

Table 5. Checklist Table. Please fill out this checklist table in the answer column.

Requirements	Answer
A meaningful title	Yes
The number of authors (≤6)	3
Author affiliations and ORCID	Yes
Corresponding author email is presented	Yes
Validation scores are presented in the abstract	Yes
Introduction includes at least three parts: background, related work, and motivation	Yes
A pipeline/network figure is provided	Fig. 1
Pre-processing	Pages 3, 4
Strategies to use the partial label	Page 5
Strategies to use the unlabeled images	Page 5
Strategies to improve model inference	Page 4
Post-processing	Pages 5, 6
Dataset and evaluation metric section is presented	Page 5
Environment setting table is provided	Table 1
Training protocol table is provided	Table 2
Ablation study	N/A
Efficiency evaluation results are provided	Table 4
Visualized segmentation example is provided	Figures 2, 3
Limitation and future work are presented	Yes
Reference format is consistent.	Yes