A Semi-Supervised Multi-Organ Segmentation Method via Cross Teaching between CNN and Transformer

Ru Wang[1] and Kunlan Xiang[2]

[1] Xi'an University of Technology, No.5 South Jinhua Road, Xi'an 710048, China
[2] University of Electronic Science and Technology of China, Chengdu, China

Abstract. The automatic segmentation of abdominal CT multi -organs can improve the efficiency of clinical work processes such as disease diagnosis, prognosis analysis, and treatment plan. However, for medical images, the acquisition of data is usually expensive, because it requires professional knowledge and time to generate accurate annotations. We proposed a cross teaching semi-supervised medical image segmentation model based on CNN and Transformer. At the same time, two deep neural networks were trained, and their mutual teaching combined their respective learning paradigms to improve model performance. The segmentation of the experiment on the data shows that our model is effective. Our experiment show that the the average running time spend on one data is 719.9262(s), maximum GPU memory required in our experiment is 269, average area under GPU memory time curve is 191794.82, and average area under GPU memory time curve is 36302.29.

Keywords: Semi-supervised learning · Medical Image Segmentation · cross teaching.

1 Introduction

The abdomen refers to the part of the chest diaphragm to the pelvis, which contains many important organs in the human body, such as the liver, right kidney, spleen, pancreas, aorta, inferior vena cava, right adrenal gland, left adrenal gland, gallbladder, esophagus, stomach, duodenum and left kidney. CT scan is a regular examination method for diagnosis of abdominal related diseases. The division of abdominal organs in the abdominal CT image accurately helps the study of the segmentation detection algorithm of organ lesions, which can help doctors formulate more accurate surgical solutions, and it is also an important step in the three-dimensional reconstruction of the abdominal organs. A large number of deep learning algorithms have obtained a good multi-organ division result, but these solutions usually have large model size and require a large number of computing resources. It is impractical to deploy in clinical practice. Therefore,it is necessary to develop a fast, low GPU memory and can use data without labels abdominal multi-organ segmentation DL architecture that fits real clinical practice and requirements in terms of both accuracy and efficiency.

CNN-based medical image segmentation approaches have been studied for many years, and most of them are based on UNet [21] or its variants, achieving very promising results in various tasks [13]. Although the exceptional representation capacity, CNN-based methods are also limited by lacking the ability of modeling the global and long-range semantic information interaction, due to the intrinsic locality of convolution operations [3]. More recently, self-attention-based architectures [7] (vision transformers) are introduced to the vision recognition tasks to model the long-range dependencies. After that, many variants of vision transformers achieved great success in natural image recognition tasks, like Swin-Transformer [16], DieT [25], PVT [28], TiT [8], etc. Benefiting from the great representation capacity of transformers, several works attempt to use transformers to replace or combine CNNs for better medical image segmentation results, such as TransUNet [3], Swin-UNet [2], CoTr [29], UNETR [9], nnFormer [32], etc. All these works show that transformers can further lead to performance gain than CNNs and also point out that it is worth to pay more attention to the transformer in the future. Although transformers have very exceptional representation capacity, it is still a data-hungry solution for recognition tasks, even require more data than CNNs [10,23].

In order to alleviate the lack of medical image labeling data, make full use of a large amount of unlabeled data, and reduce the cost of labeling, many methods have been proposed in recent years to develop a high-performance medical image segmentation model. Semi-supervised learning frameworks directly learn from limited labeled data and a large amount of unlabeled data to obtain high-quality segmentation results, which have attracted great attention in the field of medical image computing community. A lot of semi-supervised methods have been proposed for medical image analysis, including pseudo-labelling [27,1,5], deep co-training [20,33], deep adversarial learning [31], mean teacher and its extensions [24,30,15], multi-task learning [17,14,4], confidence learning [26], contrastive learning [19], and etc. All these methods combine both labeled and unlabeled data to train powerful and robust CNN models. How to train transformers with a semi-supervised fashion is also an interesting and challenging problem, especially for data limited medical image analysis tasks.We have conducted semi-supervised segmentation exploration of 13 organs, but there is no contribution work.

2 Method

2.1 Preprocessing

Our model does not use 3D patch to process data in 3D, but to handle 2D axial slices independently. The image size for axial slicing is 512 × 512.each CT volume is clipped to the [1, 99] percentiles of the intensity values. In addition, a z-score normalization is applied based on the mean and standard deviation of the intensity values among the whole training dataset. Neither cropping nor resampling is employed.

2.2 Proposed Method

In order to meet the clinical needs and meet the low GPU memory, but still effectively monitor the DL model, we have built a cross-teaching semi-supervised network architecture with two different system structures and different initialized parameter branches. The two branches in our framework are cnn based on vgg and transformer. Through cross-teaching strategies, you can use data without labels for training.

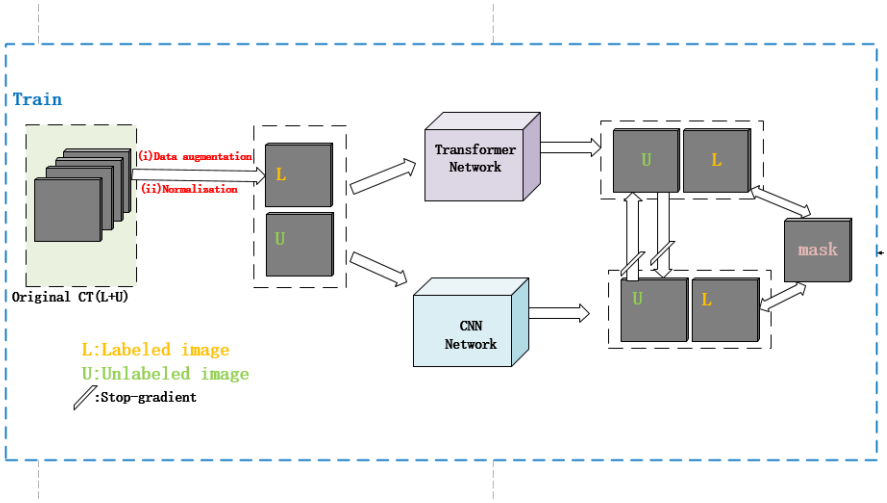

Fig. 1. Network architecture

The framework uses the processed 2D images as input, and each input image produces predictions through a CNN and a Transformer. For the general semi-supervised learning, the training set always consists of two parts:labeled data set with N annotated images and unlabeled data set D_M^u with M raw images ($M >> N$), the entire train set D_N^l For an image $x_i \in D^u$,its ground truth y_i is available. In contrast, if $x_i \in D^u$ its ground truth is not provided. In this work, the proposed Cross Teaching between CNN and Transformer is depicted in Fig. 1.If $x_i \in D^u$ a commonly-used supervised loss function is used to update models' parameters. When belongs to Du, we use a cross teaching strategy to cross supervise between a CNN $\left(f_\phi^c(.)\right)$ and a Transformer $\left(f_\phi^t(.)\right)$ for the updating of the parameters.

we introduce the perturbation in both learning paradigm-level and output-level.For an input image x_i, the proposed framework produces two predictions:

$$p_i^c = f_\phi^c\left(x_i\right); \quad p_i^t = f_\phi^t\left(x_i\right) \tag{1}$$

where p_i^c, p_i^t represent the prediction of a CNN $\left(f_\phi^c(.)\right)$ and a Transformer $\left(f_\phi^t(.)\right)$, respectively. As previously mentioned, CNN and Transformer are different learning paradigms forvision recognition, where CNN relies on the local convolution operation and the Transformer is based on the long-range self-attention, so these predictions have different properties essentially in the output level. Based on the predictions of $\left(f_\phi^c(.)\right)$ and $\left(f_\phi^t(.)\right)$, the pseudo labels for the cross teaching strategy are generated by this way:

$$pl_i^c = argmax\left(p_i^t\right) = argmax\left(f_\phi^t\left(x_i\right)\right); \quad pl_i^t = argmax\left(p_i^c\right) = argmax\left(f_\phi^c\left(x_i\right)\right) \tag{2}$$

where pl_i^c, pl_i^t are generated pseudo labels for the CNN $\left(f_\phi^c(.)\right)$ and the Transformer $\left(f_\phi^t(.)\right)$ training, respectively. It's worthy to point that pl_i^c, pl_i^t are pseudo segmentation results,and there is no gradient back-propagation between pl_i^c and pl_i^t, and between pl_i^c and pl_i^t in each mini-batch. Then, the cross teaching loss for the unlabeled data is defined as:

$$\mathcal{L}_{ctl} = \underbrace{\mathcal{L}_{dice}\left(p_i^c, pl_i^c\right)}_{supervision for CNNs} + \underbrace{\mathcal{L}_{dice}\left(p_i^t, pl_i^t\right)}_{supervision for Transformers} \tag{3}$$

where $\mathcal{L}_{dice}$ is the standard dice loss function. Differently from consistency regularization loss, the cross teaching loss is a bidirectional loss function, one stream is from the CNN to the Transformer and the other is the Transformer to the CNN, there are no explicit constraints to enforce their predictions to become similar. In our framework, the transformer is also just used for complementary training, not used to produce final predictions. The overall training objective function is a joint loss with two parts, a supervised loss on the labeled data and an unsupervised loss for the unlabeled data.The supervised loss $\mathcal{L}_{sup}$ consists of two widely-used loss functions:

$$\mathcal{L}_{sup} = \mathcal{L}_{ce}\left(p_i, y_i\right) + \mathcal{L}_{dice}\left(p_i, y_i\right) \tag{4}$$

where $\mathcal{L}_{ce}$, $\mathcal{L}_{dice}$ are the cross-entropy loss and dice loss, respectively.p_i, y_i represent the prediction and label of image x_i. The overall loss function is defined as :

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda \mathcal{L}_{ctl} \tag{5}$$

In order to avoid a large amount of GPU memory consumption, we designed a 2D multi-type division model. The network independently process axial slice to generate 2D mask, and then add these masks to a 3D zero-value body of the original size. In order to use the spatial relationship between the abdominal structure, the model learns to depict multiple organs at the same time, not depending on the specific model of multiple organs.We keep the largest connected segmented areas for voxels respectively labeled as liver, right kidney, spleen, pancreas, aorta, inferior vena cava, right adrenal gland, left adrenal gland, gallbladder, esophagus, stomach, duodenum, left kidney.No ensembling method is

used.Finally, we further change the predicted data to np.uint8 type,and we don't use any post-processing strategy.

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 [22], KiTS [11,12], AbdomenCT-1K [18], and TCIA [6]. 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 development environments and requirements are presented in Table 1.

Table 1. Development environments and requirements.

Windows/Ubuntu version	Ubuntu 20.04.4 LTS
CPU	Intel(R) Core(TM) i9-10850K CPU @ 3.60GHz 3.60 GHz
RAM	48GB
GPU (number and type)	NVIDIA GeForce RTX 3070 8G
CUDA version	11.0
Programming language	Python 3.8
Deep learning framework	Pytorch (Torch 1.7.1, torchvision 0.8.2)
Specific dependencies	
(Optional) Link to code	

Training protocols At present, the 3D segmentation effect is still less than 2D. Due to the memory limit, the whole picture cannot be trained. It can only be used in the form of patch.Compared with the 2D segmentation network, the parameters of the 3D segmentation network have increased significantly, and more parameters need more training data, so we choose the 2D training network.We use PyTorch for all method's implementations, and run all experiments on a Ubuntu desktop with a GTX3070 GPU. All these networks are trained by the SGD optimizer with a batch size of 2, where half of them are labeled in batch for semi-supervised learning. The poly learning rate strategy is used to adjust the learning rate, where the initial learning rate is set to 0.01.

Table 2. Training protocols.

Network initialization	"kaiming" and "xavier"normal initialization
Batch size	2
Patch size	4×4
Total epochs	19
Optimizer	SGD with nesterov momentum ($\mu = 0.9$)
Initial learning rate (lr)	0.01
Lr decay schedule	poly
Training time	18 hours
Number of model parameters	58.62M[3]
Number of flops	12.32GFlops [4]
CO_2eq	5.236665g [5]

4 Results and discussion

4.1 Quantitative results on validation set

The dice values are shown in Table 3.

4.2 Segmentation efficiency results

Figure2 presents a ralatively successful predicted label map for an example from the validation set. While figure3 presents a failed predicted label map for an example from the validation set.

Table 3 presents the average DSC and NSD scores for the thirteen organs. Overall, the performance of learning with full labeled images is higher than learning with the both labeled images and unlabeled images. Liver segmentation obtains the best DSC and NSD scores with compact distributions. The low scores and dispersed distributions of NSD reveal relatively high boundary errors because of the effects of various pathological changes.The overall segmentation

Table 3. The DSC comparisons between between with and without using unlabelled images.

Structure	**labelled**	**labelled+unlabelled**
liver	0.1958	0.047636
Right kidney	0.0002	0
Spleen	0	0.02
Pancreas	0.0325	0.00566
Aorta	0.0331	0.01391
IVC	0.0245	0.019132
RAG	0	0.02
LAG	0	0.04
Gallbladder	0	0.12
Esophagus	0.0122	0
Stomach	0.0347	0.000556
Duodenum	0.0429	0.020054
Left kidney	0	0.04

Table 4. Quantitative results on validation set

2*structure	DSC		NSD	
	Mean	STD	Mean	STD
Liver	0.047636	0.08028838	0.054956	0.0418234
RK	0	0	0	0
Spleen	0.02	0.14142136	0.02	0.1414214
Pancreas	0.00566	0.02084338	0.018146	0.0513465
Aorta	0.01391	0.03123522	0.01794	0.0290052
IVC	0.019132	0.05116809	0.017942	0.0370661
RAG	0.02	0.14142136	0.02	0.1414214
LAG	0.04	0.19794866	0.04	0.1979487
Gallbladder	0.12	0.32826072	0.12	0.3282607
Esophagus	0	0	0	0
Stomach	0.000556	0.00299892	0.002764	0.0111379
Duodenum	0.020054	0.05779298	0.06178	0.1308669
LK	0.04	0.19794866	0.04	0.1979487

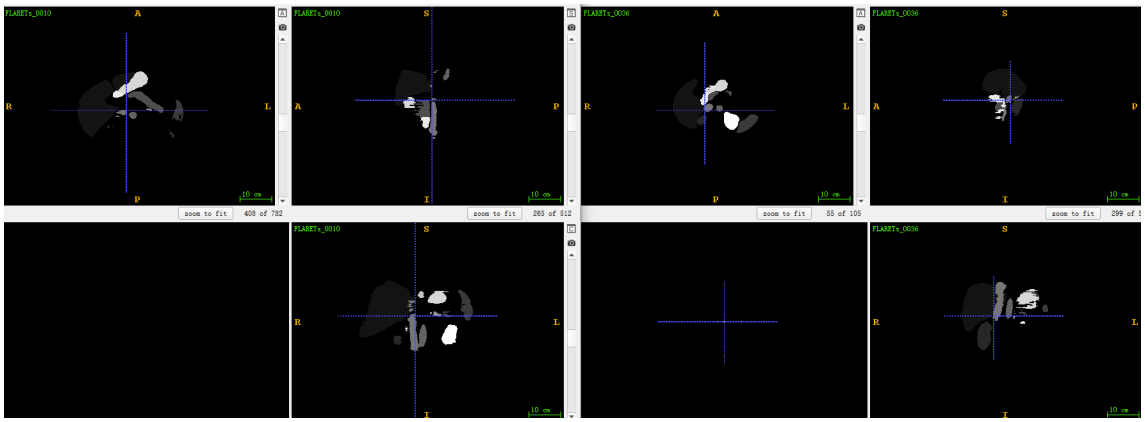

Fig. 2. two relatively successful predicted label maps

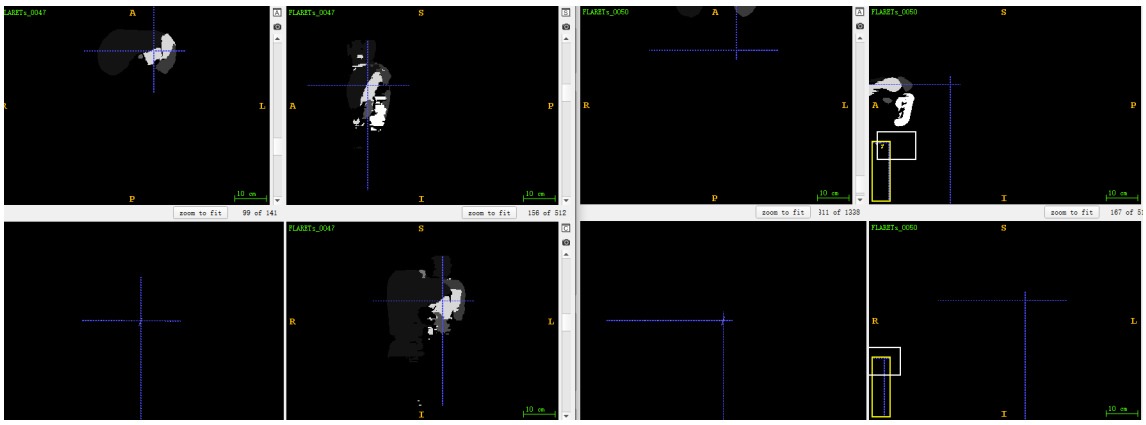

Fig. 3. two failed predicted label maps

evaluation index values are very low. Looking at the segmented image, it can be seen that the organs are separated, but the position of the organs have changed, so the evaluation index values are very low.

4.3 Limitation and future work

Three of the 13 organs have not been learned well by the network. In the future, it will further improve the network and improve the learning ability of the network, so that it can well divide all 13 organs.At present, although the result of the segmentation does seem to be the organ that needs to be segmented, the positions of the organs have shifted, which may be caused by the translation invariance of the network, the limited learning ability of the network, or inappropriate data augmentation.In the future, we will further explore suitable data enhancement methods, further analyze the characteristics of the dataset, and construct a more effective segmentation network.

5 Conclusion

The segmentation network of this article can effectively use without label data. However, the current segmentation effect of the network is not good, and the location of the segmentation organs has shifted, which needs to continue to explore how to build a more reasonable segmentation network structure.The separation network of this article can divide most organs, but the location of the organs has shifted, and the cause of exploration of offset needs to continue to improve. The method of semi-supervision can effectively use unlabeled data. If the divided organs can not be offset, it will obtain a segmentation network that can make full use of the unlabeled data.

Acknowledgements We has not used any pre-trained models nor additional datasets other than those provided by the organizers in this competition,and our proposed solution is fully automatic without any manual intervention.

References

1. Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P.M., Rueckert, D.: Semi-supervised learning for network-based cardiac mr image segmentation. In: MICCAI. pp. 253–260. Springer (2017) 2
2. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537 (2021) 2
3. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021) 2

4. Chen, S., Bortsova, G., Juárez, A.G.U., van Tulder, G., de Bruijne, M.: Multi-task attention-based semi-supervised learning for medical image segmentation. In: MICCAI. pp. 457–465. Springer (2019) 2
5. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2613–2622 (2021) 2
6. 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) 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) 2
8. Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y.: Transformer in transformer. arXiv preprint arXiv:2103.00112 (2021) 2
9. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H., Xu, D.: Unetr: Transformers for 3d medical image segmentation. arXiv preprint arXiv:2103.10504 (2021) 2
10. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377 (2021) 2
11. 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) 5
12. 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
13. 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) 2
14. Kervadec, H., Dolz, J., Granger, E., Ayed, I.B.: Curriculum semi-supervised segmentation. In: MICCAI. pp. 568–576. Springer (2019) 2
15. Li, X., Yu, L., Chen, H., Fu, C.W., Xing, L., Heng, P.A.: Transformation-consistent self-ensembling model for semisupervised medical image segmentation. TNNLS **32**(2), 523–534 (2020) 2
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021) 2
17. Luo, X., Chen, J., Song, T., Wang, G.: Semi-supervised medical image segmentation through dual-task consistency. AAAI **35**(10), 8801–8809 (2021) 2
18. 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 5
19. Peng, J., Wang, P., Desrosiers, C., Pedersoli, M.: Self-paced contrastive learning for semi-supervised medical image segmentation with meta-labels. In: Thirty-Fifth Conference on Neural Information Processing Systems (2021) 2

20. Qiao, S., Shen, W., Zhang, Z., Wang, B., Yuille, A.: Deep co-training for semi-supervised image recognition. In: ECCV. pp. 135–152 (2018) 2
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241 (2015) 2
22. 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) 5
23. Tang, Y., Yang, D., Li, W., Roth, H., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis (2021) 2
24. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NeurIPS. pp. 1195–1204 (2017) 2
25. 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
26. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: CVPR. pp. 2517–2526 (2019) 2
27. Wang, G., Zhai, S., Lasio, G., Zhang, B., Yi, B., Chen, S., Macvittie, T.J., Metaxas, D., Zhou, J., Zhang, S.: Semi-supervised segmentation of radiation-induced pulmonary fibrosis from lung ct scans with multi-scale guided dense attention. IEEE Transactions on Medical Imaging (2021) 2
28. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. arXiv preprint arXiv:2102.12122 (2021) 2
29. Xie, Y., Zhang, J., Shen, C., Xia, Y.: Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. arXiv preprint arXiv:2103.03024 (2021) 2
30. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: MICCAI. pp. 605–613. Springer (2019) 2
31. Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z.: Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: MICCAI. pp. 408–416. Springer (2017) 2
32. Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y.: nnformer: Interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201 (2021) 2
33. Zhou, Y., Wang, Y., Tang, P., Bai, S., Shen, W., Fishman, E., Yuille, A.: Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training. In: WACV. pp. 121–140. IEEE (2019) 2