Prompt-based Segmentation Model of Anatomical Structures and Lesions in CT Images

Xi Ouyang[1], Dongdong Gu[1], Xuejian Li[1], Wenqi Zhou[1,2], Qianqian Chen[1,2], Yiqiang Zhan[1], Xiang Zhou[1], Feng Shi[1], Zhong Xue[3](✉), and Dinggang Shen[1,2,4](✉)

[1] Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China
[2] School of Biomedical Engineering, ShanghaiTech University, Shanghai, China
[3] Shanghai United Imaging Intelligence Co., Ltd., Beijing, China
[4] Shanghai Clinical Research and Trial Center, Shanghai, China
zhong.xue@ieee.org, dgshen@shanghaitech.edu.cn

Abstract. Deep learning models have been successfully developed for various medical image segmentation tasks. However, individual models are commonly developed using specific data along with a substantial amount of annotations, ignoring the internal connections between different tasks. To overcome this limitation, we integrate such a multi-task processing into a general computerized tomography (CT) image segmentation model trained on large-scale data, capable of performing a wide range of segmentation tasks. The rationale is that different segmentation tasks are often correlated, and their joint learning could potentially improve overall segmentation performance. Specifically, the proposed model is designed with a transformer-based encoder-decoder architecture coupled with automatic pathway (AP) modules. It provides a common image encoding and an automatic task-driven decoding pathway for performing different segmentation tasks via specific prompts. As a unified model capable of handling multiple tasks, our model not only improves the performance of seen tasks but also quickly adapts to new unseen tasks with a relatively small number of training samples while maintaining reasonable performance. Furthermore, the modular design of automatic pathway routing allows for parameter pruning for network size reduction during the deployment.

Keywords: Foundation model · CT image segmentation · Large vision model.

1 Introduction

Medical artificial intelligence (AI) applications can help physicians expedite and make more precise assessments and choose better treatment options [21,24]. In radiation therapy, accurate segmentation of tumors and organs at risk in medical images is essential for dose planning and delivery [22]. Thus, image segmentation is considered fundamental for medical AI development.

Considering the diversity of medical images and variability of organ sizes and shapes, the majority of image segmentation jobs [6,25,18] are usually carried out in a data- and task-specific manner to achieve plausible performance. Specific neural network models are tailored to particular segmentation tasks, and these models are usually unable to capture the relationship between different tasks, and their performance may drop significantly when the input image data and/or the target segmentation deviate from those employed during training. It is also difficult to transfer the task-specific models to new unseen tasks. In the literature, some studies adopt the multi-task learning (MTL) strategy [5,26], which involves appending multiple heads to the end of a shared neural network model to generate predictions for different tasks. However, multi-head frameworks usually require a large number of parameters in both the encoder and decoder to achieve high performance across tasks, which poses challenges when deploying them in clinical applications. This is because almost all the parameters (excluding those in the final heads) are necessary even for a single segmentation task. Recently, the segment anything model (SAM) [11] has drawn considerable attention as a powerful pre-trained network. However, there exists a significant distribution shift from natural images to medical images, which renders direct application in most medical contexts challenging due to the huge gap between 3D or 4D medical images and 2D natural images. Moreover, the use of SAM typically necessitates guidance through appropriate location clues (foreground points, boxes, or masks) to acquire desired segmentation results, also, the quality of the results varies with the different cues provided, hindering SAM from achieving a fully automated process in CAD systems.

In this paper, we integrate multi-task segmentation using a prompt-based CT image segmentation model, which is trained on a large number of multi-task datasets as a foundation model for CT image segmentation. Unlike common task-specific AI models, our model is pre-trained on large-scale multi-task datasets and capable of feature extraction with a transformer-based encoder across various tasks. Compared to multi-head-based MTL, our model utilizes comprehensive automatic pathway (AP) routing to automatically determine the feature decoding routes for different tasks. This allows our model to capture correlations between different tasks while enhancing the segmentation performance of each. This idea is similar to the network architecture search (NAS) method [3] to automate the discovery of optimal network structures. With AP routing, we can mitigate conflicts and maintain mutual benefits across various tasks.

We built up the most extensive dataset so far for training the model, which contains 32,170 cases of CT scans with 58,499 annotations for 83 tasks. To the best of our knowledge, this is the largest dataset consisting of both non-contrast CT and contrast-enhanced CT so far for medical image segmentation training. This large-scale data training approach facilitates the model to learn transferable and generalizable representations in medical images, enabling it to migrate to other new medical image tasks. Experimental results on 15 tasks demonstrate that our method outperforms the state-of-the-art methods on various tasks.

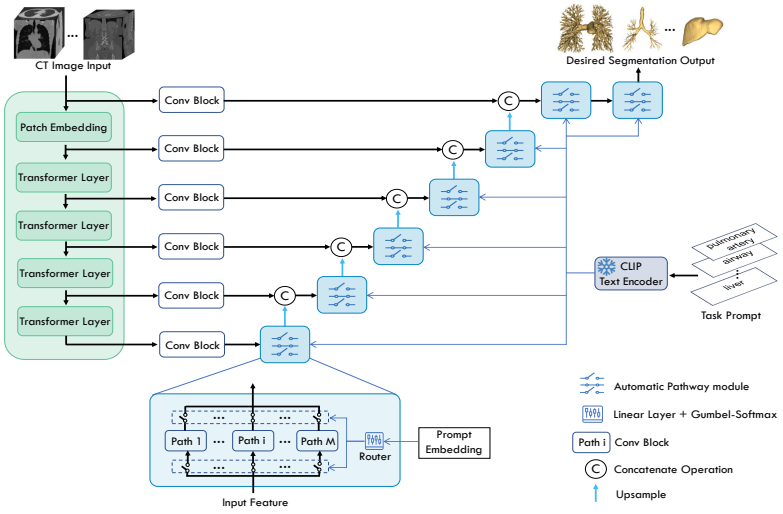

Fig. 1. The overall structure of our model. The decoder encompasses multiple layers with automatic pathway (AP) modules using task-specific prompts as input.

2 Method

As illustrated in Fig. 1, the proposed model consists of three components: 1) A transformer-based encoder that extracts features from various medical images and forms a common interactive latent space; 2) a prompt text encoder that provides prompt embedding to guide the sub-pathway selection in the AP modules; and 3) a decoder that processes data through different sub-pathways to generate the corresponding segmentation results based on automatic routing using the AP modules. We outline the respective details below.

2.1 Overall Structure

Given the remarkable performance of the Swin transformer [13], we extend its structure into 3D as [7] and adopt the encoder in our model, which consists of four stages, and each stage contains two transformer blocks using regular-window-partitioning multi-head self-attention modules. The input of the Swin-transformer encoder is a patch embedding layer with a patch size of $2 \times 2 \times 2$ and 48 output channels. Five-level downsampling operations with $1/2$ scale are used in the encoder, which has a size of 8.06 million parameters in total to learn the image information from the large-scale dataset.

The decoder is designed based on 3D residual blocks at different levels, and skip connections are used similarly to the SwinUnetr [7]. For the sake of flexibility and adaptability in multi-task segmentation, we integrate a sequence of automatic pathway (AP) modules within the decoding pipeline. These AP modules enable automatic and independent routing of different sub-pathways at each

decoding level, thus convolutional blocks within each sub-pathway are automatically activated and executed depending on the purpose/prompt fed to our model.

Specifically, we use the text prompt as guidance to direct the model about which object to be segmented, *e.g.*, "liver" indicates liver segmentation in CT images. To encode the prompt information into our model, we use the text encoder from the CLIP model [19] to transfer the input prompt into a latent vector representation, called prompt embedding. The reason to use the existing pre-trained model is that it provides promising projections and gives reasonable initialization for different tasks. In our experiments, the parameters of this text encoder are fixed during the training to maintain the ability to understand texts from the large-scale pre-training of CLIP. Details of AP module are described in the next section.

2.2 Automatic Pathway Routing

The objective of the AP module is to enhance the model's ability to capture correlations and reduce conflicts between various tasks. By training on the large-scale, multi-task dataset, this module aims to improve the segmentation performance of each task while ensuring mutual benefits across tasks. The core mechanism of AP modules is to select suitable sub-pathways at different levels given a prompt embedding vector. Specifically, an AP module consists of two essential components: a routing layer and M candidate sub-pathways. The routing layer is formed by fully connected layers and a Gumbel softmax layer to learn a transformation between prompt embedding and sub-pathway routing. Each sub-pathway is a convolutional network module in the decoding process. More precisely, the text/prompt embedding features of a given task, derived from the pre-trained text encoder, are utilized for the automatic sub-pathway selection using routing layers and Gumbel softmax. The M candidate sub-pathways have the same network structure (3D residual blocks) but different parameters after training. This way, different routes of image feature processing are automatically used for different tasks at various levels.

The proposed AP module enables automatic and independent routing of different sub-pathways at each decoding level. With M candidate sub-pathways at each decoding level, the AP-enabled decoder can be configured by a combination of sub-pathway selections at different levels, making it suitable for handling a large number of tasks.

3 Experiments

3.1 Datasets

We collected the largest dataset for medical image segmentation tasks to date for the network training, which contains 32,170 CT scans with 58,499 annotations corresponding to 83 segmentation tasks throughout the entire body. Among the datasets, a total of 45,725 annotations of in-house CT scans were collected

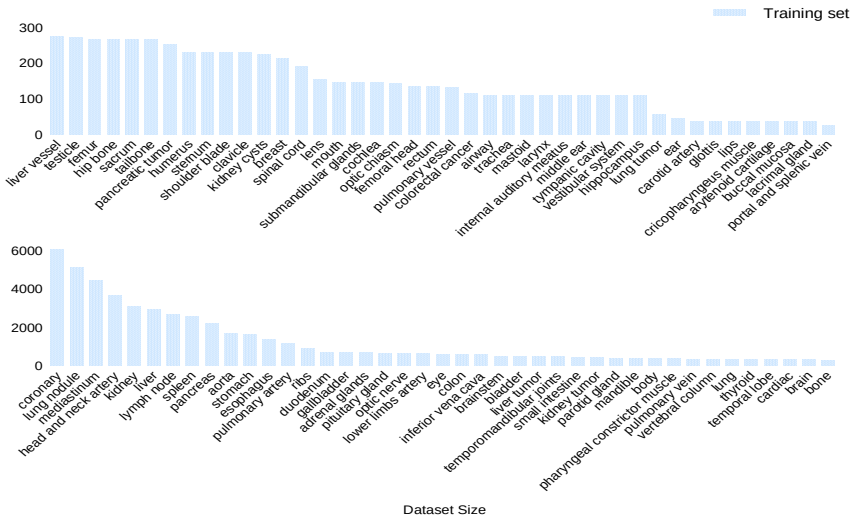

Fig. 2. Overview of the datasets. The datasets cover a wide range of medical image segmentation tasks that span across the entire body. Overall, the training set encompasses 58,499 annotations.

from our collaborating hospitals, and IRB approvals were obtained by the Research Ethics Committees from those centers. Written informed consent was waived because of the retrospective nature of the study. In addition, the rest 12,774 annotations in the experiments came from multiple publicly available datasets, *i.e.*, from the Head and Neck organ-at-risk CT & MR Segmentation dataset (HaN-Seg) [17], The Cancer Imaging Archive (TCIA) [4], the Medical Segmentation Decathlon (MSD) [1], the Segmentation of Organs-at-Risk and Gross Tumor Volume of NPC for Radiotherapy Planning Challenge (SegRap, https://segrap2023.grand-challenge.org/), the Multi-Atlas Labeling Beyond the Cranial Vault (BTCV) [12], the Liver Tumor Segmentation (LiTS) Challenge [2], the Kidney Tumor Segmentation (KiTS) Challenge [8], the Whole abdominal ORgan Dataset (WORD) [15], the CT-ORG dataset [20], and the Multi-Modality Abdominal Multi-Organ Segmentation (AMOS) Challenge [10]. More details about the datasets can be found in their corresponding references.

3.2 Experimental Settings

The networks are implemented using the PyTorch framework [16], with memory usage and computation speed optimized using the automatic mixed precision (AMP) package. The training process employs the AdamW optimizer [14] with a momentum of 0.9, weight decay of 0.00001, and an initial learning rate of 0.0002. For data augmentation, common strategies for deep-learning model training are employed, including rotation, scaling, flipping, shifting, and adding noise. For

Table 1. Quantitative comparison of Dice scores (%) for segmentation of 7 non-tubular-structure organs, and 1 tumor. The sole-path model represents our model without the automatic pathway module trained in each task. We also show the number of training and testing cases ("training/testing") after the task name. * indicates a significant difference compared to our proposed method (p-value < 0.05). Standard deviation (std) results are shown in Table A.1 of supplementary.

Method	Stomach 1,626/183	Lung 345/20	Ear 45/5	Bladder 501/57	Sacrum 266/35	Cardiac 330/20	Thyroid 334/38	Lung tumor 57/6
nnU-Net	90.91	**98.80***	86.05	90.79	92.57	87.86*	**89.23**	77.12
Cascaded Vb-Net	91.59	98.14	80.99	86.13*	90.54	91.26	83.00*	59.11
Sole-path model	90.45*	98.40	83.93	83.13*	91.16	89.62	85.15*	76.40
Ours	**92.59**	98.61	**86.29**	**91.68**	**94.86**	**93.01**	88.90	**80.75**

the adjustment strategy of the learning rate, the model is trained with cosine annealing with the WarmRestarts strategy, and the number of warmup epochs is set to 50 in the experiments.

Totally 20 NVIDIA A40 graphics processing units are used for training. First, the model is trained for 3,000 epochs with all the training data of 83 tasks. Three candidate spacing settings, *i.e.*, $0.6 \times 0.6 \times 0.6mm^3$, $1.0 \times 1.0 \times 1.0mm^3$ and $1.5 \times 1.5 \times 1.5mm^3$, are used. During training, it has a 10% chance to randomly choose among three candidate spacing settings, and 90% chance to choose the configuration closest to the original image spacing. Subsequently, we normalize the processed intensity values in the CT images to the range [0, 1] following "window width/level" (window: 1800, level: 0) operation. In our experiments, we set M as 6 in all the AP modules. Second, for each selected sub-pathway in AP modules, we only select the data of the tasks belonging to this sub-pathway to further update the parameters for another 1,000 epochs of the corresponding sub-pathway with other parameters in the model are fixed.

3.3 Evaluation Metrics

The performance is mainly evaluated by two commonly used metrics Dice and clDice. Dice provides a globally averaged assessment of segmentation performance across all voxels, and clDice evaluates the topological degree of matching for tubular-structure tasks. Also, we show the values of Topology Precision (Tprec) and Topology Sensitivity (Tsens) [23], detailing the sensitivity and precision for segmentation results. The two-tailed t-test is used to compare the results with those obtained by gCIS.

3.4 Experimental Results

Performance on segmentation of non-tubular structures. To better demonstrate the effectiveness of our model, we compare the proposed foundation model with the state-of-the-art (SOTA) models, *i.e.*, nnU-Net [9] and cascaded Vb-Net

Table 2. Quantitative comparison of Dice scores (%) and clDice scores (%) for segmentation of 7 tubular-structure organs, including the number of training and testing cases. * indicates a significant difference compared to our proposed method (p-value < 0.05). Std results are shown in Table A.2 of supplementary.

Metric	Method	Airway 108/20	Coronary 6,065/155	Pulmonary artery 1,153/20	Pulmonary vein 352/20	Aorta 1,694/20	Lower limb artery 634/73	Head & neck artery 3,664/729
Dice (%)	nnU-Net	87.27	75.85*	88.17	87.02	**95.49**	76.74*	87.80*
	Cascaded Vb-Net	85.70*	**83.02**	87.50	85.70	89.54*	79.73	93.19*
	Sole-path model	86.79	79.55*	87.71	84.07	90.90*	78.86	93.50*
	Ours	**87.60**	82.72	**88.83**	**87.16**	95.31	**80.13**	**94.45**
clDice (%)	nnU-Net	82.96	84.43*	87.32	90.29	99.83	79.23	81.64*
	Cascaded Vb-Net	78.49*	88.10*	85.88*	87.59*	96.46*	75.80*	91.29*
	Sole-path model	83.10	86.68*	85.86*	88.38*	94.92*	78.37*	91.70*
	Ours	**84.47**	**90.72**	**88.58**	**92.08**	**99.88**	**82.38**	**93.28**

[22], each independently trained for individual tasks. For ablation study, we conduct comparative experiments by training sole-path models (without AP modules) separately for each task. In Table 1, we show the Dice scores for segmenting 7 non-tubular-structure organs, and 1 tumor. Standard deviation (std) results are shown in Table A.1 of supplementary. It can be observed that our method can yield the highest Dice scores for 6 tasks, which proves its superior performance. Moreover, most std of Dice of our model are lower than other methods, which proves that our model is more robust. For thyroid segmentation, our model performs slightly less favorably than nnU-Net, which could be attributed to the nature of the testing set, which comprises exclusively thick CT images (3mm), rendering the precise delineation of thyroid boundaries challenging. Meanwhile, our model has a smaller std, indicating that our model has a consistent and stable performance for all cases in testing set.

For the segmentation of "lung tumor", our model can achieve the Dice score of 80.75% while nnU-Net model can only reach 77.12%. It is a challenging segmentation task due to two key factors: 1) the limited training dataset consisting of only 57 cases, and 2) the considerable diversity of tumor size and location. In this task, our model could improve the Dice score by approximately 5% compared with the nnU-Net baseline. Overall, our model is more robust for challenging tasks and is more friendly to data-limited tasks.

Performance on segmentation of tubular structures. Meanwhile, we compare Dice and clDice scores with state-of-the-art methods for the segmentation tasks of 7 tubular structures in Table 2. The tasks include segmentation of the coronary, pulmonary artery and vein, aorta, lower limb artery, head and neck artery, as well as airway segmentation. Concurrently, it is worth noting that the

Dice score tends to favor the volumetric segmentation of larger tubular components, while clDice could better reveal the connectedness of tubular structure. It can be observed that our model consistently outperforms other state-of-the-art methods in terms of clDice scores in all tasks, showing that it is a power model for these challenging tubular targets. For the segmentation of the lower limb artery, our method outperforms the SOTA results by nearly 3% in clDice scores although the coverage range of lower limb arteries is large and the shape is complex.

Performance of few-shot learning on a new task. It is important to rapidly develop AI models for new tasks in limited annotated data in real clinical scenarios. In this section, we show the great generalization ability of our foundation segmentation model for the segmentation of the renal artery. We collected a total of 33 cases for this new task, and it is worth noticing that neither these images nor these annotations have been used in the multi-task co-training procedure of our model. 28 out of these cases are randomly selected into the testing set. As shown in Table 3, we use two settings to train all the methods, *i.e.*, 1-shot and 5-shot. The n represents the number of training cases in the n-shot setting.

The method for rapidly adapting our model to new tasks consists of three steps. The first step is to use text prompts to select suitable sub-pathways in AP modules. We use “renal artery" to choose the sub-pathways for this new task. Then, we keep the selected sub-pathways and trim the remaining sub-pathways from the network to reduce the size of our network. Since there are originally 6 candidate sub-pathways in each AP module, the network trimming operation can reduce the parameters of AP modules to 1/6. In contrast, the original foundation model for all tasks has 60.50 million parameters, while the trimmed model for this new task has only 26.75 million parameters. It could be a great advantage to deploy the proposed model for downstream tasks in real clinical centers. Finally, during the training procedure, the parameters in the encoder of our model are fixed, while only the parameters of the decoder with these specific sub-pathways are updated. The fixed encoder offers a powerful encoding ability for CT images due to the pre-training from the large-scale dataset. At the same time, the selection of the sub-pathways by the task prompt can provide a valuable starting point for the decoder training.

We show the comparison results with nnU-Net and cascaded Vb-Net for the segmentation of renal artery using different training examples in Table 3. It can be observed that our foundation model can achieve significantly better clDice scores in the 1-shot and 5-shot settings. We can see the gaps of the proposed method with the best clDice scores from other SOTA methods are larger when the smaller number of samples in training. It proves that the fixed encoder offers a powerful encoding ability for CT images due to the pre-training from the large-scale dataset. At the same time, the selection of the sub-pathways by the task prompt can provide a valuable starting point for the decoder training.

Table 3. Comparison with other methods for the few-shot segmentation of renal artery. * indicates a significant difference compared to our proposed method (p-value < 0.05). Std results are shown in Table A.3 of supplementary.

Training Setting	Method	Dice (%)	clDice (%)	Tsens (%)	Tprec (%)
1-shot	nnU-Net	82.56	56.17*	41.88*	**97.45***
	Cascaded Vb-Net	71.53*	45.59*	27.99*	87.12
	Ours	**84.84**	**67.04**	**55.07**	87.30
5-shot	nnU-Net	**89.81**	77.40	65.45*	**98.47***
	Cascaded Vb-Net	85.79*	61.57*	47.13*	96.11
	Ours	88.64	**80.77**	**71.43**	94.30

4 Conclusion

By incorporating dozens of tasks and tens of thousands of CT volumes, we present a general image segmentation model by using an AP module-based decoder. To address the inherent constraints associated with a shared decoder for multi-head networks in traditional multi-task learning, our method enhances task flexibility through the routing of multiple sub-pathways at each decoder level, using AP modules. These sub-pathways are dynamically and autonomously selected and learned to alleviate conflicts and maximize benefits across a spectrum of tasks. Our method could achieve the best performance in 15 tubular or non-tubular structure segmentation tasks. Moreover, our foundation model can achieve excellent performance with quite limited training samples.

Acknowledgments. This research is supported by the Beijing Natural Science Foundation (IS24053).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. 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)
2. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al.: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023)
3. Chang, J., Guo, Y., Meng, G., Xiang, S., Pan, C., et al.: Data: Differentiable architecture approximation. Advances in Neural Information Processing Systems **32** (2019)
4. 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**, 1045–1057 (2013)

5. Guo, J., Liu, X., Chen, Y., Zhang, S., Tao, G., Yu, H., Zhu, H., Lei, W., Li, H., Wang, N.: Aanet: artery-aware network for pulmonary embolism detection in ctpa images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 473–483. Springer (2022)
6. Hao, D., Ding, S., Qiu, L., Lv, Y., Fei, B., Zhu, Y., Qin, B.: Sequential vessel segmentation via deep channel attention network. Neural Networks **128**, 172–187 (2020)
7. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop. pp. 272–284. Springer (2021)
8. Heller, N., Isensee, F., Trofimova, D., Tejpaul, R., Zhao, Z., Chen, H., Wang, L., Golts, A., Khapun, D., Shats, D., Shoshan, Y., Gilboa-Solomon, F., George, Y., Yang, X., Zhang, J., Zhang, J., Xia, Y., Wu, M., Liu, Z., Walczak, E., McSweeney, S., Vasdev, R., Hornung, C., Solaiman, R., Schoephoerster, J., Abernathy, B., Wu, D., Abdulkadir, S., Byun, B., Spriggs, J., Struyk, G., Austin, A., Simpson, B., Hagstrom, M., Virnig, S., French, J., Venkatesh, N., Chan, S., Moore, K., Jacobsen, A., Austin, S., Austin, M., Regmi, S., Papanikolopoulos, N., Weight, C.: The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct (2023)
9. 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)
10. Ji, Y., Bai, H., GE, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., Luo, P.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 36722–36732. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/ee604e1bedbd069d9fc9328b7b9584be-Paper-Datasets_and_Benchmarks.pdf
11. 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)
12. 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)
13. 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)
14. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
15. Luo, X., Liao, W., Xiao, J., Chen, J., Song, T., Zhang, X., Li, K., Metaxas, D.N., Wang, G., Zhang, S.: Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. Medical Image Analysis **82**, 102642 (2022)
16. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)

17. Podobnik, G., Strojan, P., Peterlin, P., Ibragimov, B., Vrtovec, T.: Han-seg: The head and neck organ-at-risk ct and mr segmentation dataset. Medical physics **50**(3), 1917–1927 (2023)
18. Qin, Y., Zheng, H., Gu, Y., Huang, X., Yang, J., Wang, L., Zhu, Y.M.: Learning bronchiole-sensitive airway segmentation cnns by feature recalibration and attention distillation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 221–231. Springer (2020)
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
20. Rister, B., Yi, D., Shivakumar, K., Nobashi, T., Rubin, D.L.: Ct-org, a new dataset for multiple organ segmentation in computed tomography. Scientific Data **7**(1), 381 (2020)
21. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Annual review of biomedical engineering **19**, 221–248 (2017)
22. Shi, F., Hu, W., Wu, J., Han, M., Wang, J., Zhang, W., Zhou, Q., Zhou, J., Wei, Y., Shao, Y., et al.: Deep learning empowered volume delineation of whole-body organs-at-risk for accelerated radiotherapy. Nature Communications **13**(1), 6566 (2022)
23. Shit, S., Paetzold, J.C., Sekuboyina, A., Ezhov, I., Unger, A., Zhylka, A., Pluim, J.P., Bauer, U., Menze, B.H.: cldice-a novel topology-preserving loss function for tubular structure segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16560–16569 (2021)
24. Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nature Biomedical Engineering **6**(12), 1399–1406 (2022)
25. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: Transbts: Multimodal brain tumor segmentation using transformer. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 109–119. Springer (2021)
26. Xu, M., Wang, Y., Chi, Y., Hua, X.: Training liver vessel segmentation deep neural networks on noisy labels from contrast ct imaging. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). pp. 1552–1555. IEEE (2020)