SegVol: Universal and Interactive Volumetric Medical Image Segmentation

Yuxin Du[1], Fan Bai[1,2], Tiejun Huang[1,3], Bo Zhao[1†]

[1]Beijing Academy of Artificial Intelligence.
[2]The Chinese University of Hong Kong.
[3]Peking University.

[†]Corresponding author: Bo Zhao <zhaobo@baai.ac.cn>

Abstract

Precise image segmentation provides clinical study with meaningful and well-structured information. Despite the remarkable progress achieved in medical image segmentation, there is still an absence of foundation segmentation model that can segment a wide range of anatomical categories with easy user interaction. In this paper, we propose a universal and interactive volumetric medical image segmentation model, named *SegVol*. By training on 90k unlabeled Computed Tomography (CT) volumes and 6k labeled CTs, this foundation model supports the segmentation of over 200 anatomical categories using semantic and spatial prompts. Extensive experiments verify that SegVol outperforms the state of the art by a large margin on multiple segmentation benchmarks. Notably, on three challenging lesion datasets, our method achieves around 20% higher Dice score than nnU-Net. The model and data are publicly available at: https://github.com/BAAI-DCAI/SegVol.

1 Introduction

Volumetric image segmentation plays a crucial role in medical image analysis by accurately extracting regions of interest, such as organs, lesions and tissues, and benefits numerous clinical applications including tumors monitoring[1], surgical planning[2], disease diagnosis[3], therapy optimization[4], etc. The research of volumetric medical image segmentation has garnered substantial attention, leading to a series of advancements[5–10].

Though the remarkable progress, existing solutions still have several key limitations that prevent their application in challenging tasks, e.g., liver tumor or colon cancer segmentation, and real-world tasks, e.g., interactive segmentation. Firstly, the publicly available volumetric medical image datasets usually consist of a small number of mask annotations from varying categories. Due to the different label spaces, the traditional segmentation models trained on one dataset have difficulty in generalizing to others. For example, the CT-ORG dataset[11–14] contains 'lungs' category, while it is 'left lung' and 'right lung' in the LUNA16 dataset[15]. The main reason is that these models do not understand the semantics of anatomical categories. Secondly, traditional segmentation models have inferior performance when segmenting complex structures, such as tumors and cysts[16]. This is because these models are trained on insufficient data and also not able to leverage the spatial information through user interaction. Last but not least, previous solutions are computationally expensive in inference process. They typically employ a sliding window to infer the whole volumetric input. This strategy is not only time-consuming but also short-sighted, as the sliding window contains only local information.

To overcome the above limitations, we introduce SegVol, a universal and interactive volumetric medical image segmentation model. It is a foundation model designed to segment more than 200 anatomical categories, delivering precise segmentation for organs, tissues, and lesions. SegVol is built on a lightweight architecture, ensuring its efficiency for practical medical image analysis. We summarize the key features of SegVol as follows:

1. Pre-train the model on 96k CTs and leverage pseudo label to decouple the spurious correlation between datasets and segmentation categories.
2. Enable text-prompt segmentation by integrating the language model into segmentation model and training it on over 200 anatomical categories of 25 datasets.
3. Employ a synergistic strategy to coordinate the semantic prompt and spatial prompt, and achieve high-precision segmentation.
4. Design a zoom-out-zoom-in mechanism that significantly reduces the computational cost, while preserving precise segmentation.

We extensively evaluate the proposed SegVol on multiple segmentation datasets. The major study involves experiments on important anatomical categories, which demonstrates the universal segmentation ability with an average Dice score of 83.02%. We compare SegVol against four state-of-the-art methods on five popular datasets, revealing its substantial superiority, especially in hard categories, where its Dice score is 14.76% higher than that of the most popular traditional method, nnU-Net[9]. Furthermore, an in-depth analysis of model's lesion segmentation ability is carried out on three tumor targets, highlighting its outstanding performance that exceeds nnU-net by 19.58% on average Dice score. Finally, we conduct ablation studies to verify the effectiveness of the zoom-in-zoom-out mechanism.

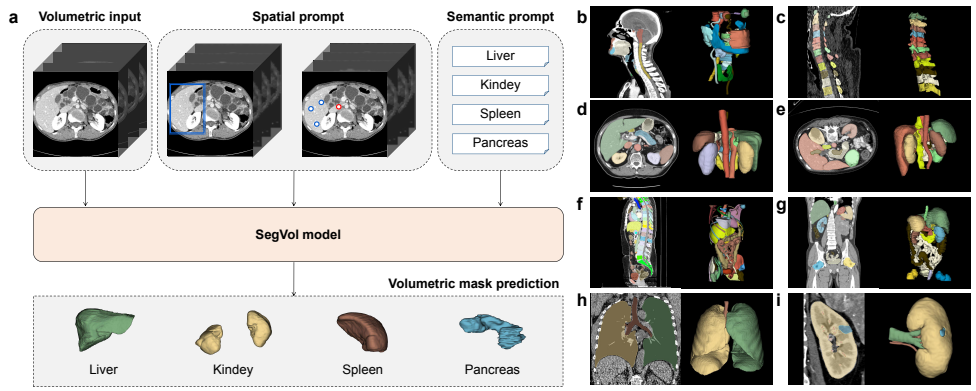

Fig. 1 An overview of SegVol and samples of joint dataset. **a** The basic input and output types of SegVol. SegVol accepts volumetric medical image, spatial prompts like bounding box prompts and point prompts and semantic prompts as inputs, and generates the corresponding volumetric mask predictions. **b-i** Samples of joint dataset for training SegVol. **b** is a head CT from HaN-Seg[17]. **c** is a vertebra CT from VerSe20[18–20]. **d-g** are abdomen CTs from AbdomenCT-12organ[21, 22], BTCV[23], TotalSegmentator[24] and WORD[25]. **h-i** are organ specific CTs from LUNA16[15] and KiPA22[26–29].

2 Results

It is a long-standing challenge in medical image analysis to built a segmentation model that is capable of handling a wide range of segmentation tasks while achieving precise segmentation results. We propose SegVol, a universal and interactive model for volumetric medical image segmentation (Fig. 1). As a universal model, our method delivers accurate segmentation results for over 200 important organs, tissues, and lesions through 'text' prompts. As a precise segmentation model, it also introduces 'point' and 'box' spatial prompts to guide the segmentation of anatomical structures, resulting in high-precision segmentation responses. To establish SegVol, we collect 6k CTs annotated with 150k segmentation masks from 25 open-source medical datasets Additionally, we crawl 90k unlabeled CT data from the web and obtain 511k pseudo volumetric masks generated by the FH algorithm[30]. We conduct systematic and comprehensive experiments on the large-scale joint dataset to train and extensively evaluate the SegVol model.

2.1 Dataset Exploration

To establish a universal volumetric segmentation model proficient in multiple tasks, we collect 25 segmentation CT datasets from medical open-source datasets to form a joint dataset, encompassing various hot issues in CT image segmentation. As shown in Figure 2 **a**, the collected joint dataset includes four main human body regions: Head and Neck, Thorax, Abdomen, and Pelvis, comprising over 200 categories of organs, tissues, and lesion types across 47 important regions. A total of 5772 CTs participate in the training and testing of the joint dataset, with a combined total of 149,199

Table 1 Primary evaluation of Dice score(%) on 19 key targets.

	Single-type prompt			Composite-type prompt	
Key targets	Text	Point	Box	Point+Text	Box+Text
Liver	95.69	84.89	93.75	95.50	96.13
Spleen	93.77	89.65	94.38	93.17	94.98
Hip	70.22	78.08	89.33	92.26	93.31
Lung	87.76	78.49	87.70	85.51	91.68
Aorta	89.88	86.41	90.33	90.19	90.83
Kidney	86.09	82.67	88.73	87.91	89.91
Trachea	88.59	87.28	88.28	88.42	89.54
Stomach	87.10	78.31	86.86	84.96	88.89
Heart	65.95	76.92	85.26	79.51	88.13
Clavicula	81.94	79.28	80.52	80.10	87.05
Scapula	83.24	75.83	80.99	82.78	84.46
Pancreas	76.48	70.85	77.38	73.34	78.70
Humerus	64.85	69.68	76.23	70.23	77.63
Colon	72.78	57.24	65.87	70.03	77.12
Vertebrae	69.07	63.80	68.61	66.41	74.11
Esophagus	69.42	66.13	70.90	66.21	71.10
Rib	66.07	66.44	59.26	70.25	70.96
Duodenum	59.65	59.41	64.94	62.44	66.88
Adrenal	57.77	48.44	64.81	48.23	66.01
Average	**77.17**	**73.67**	**79.69**	**78.29**	**83.02**

volumetric mask labels with semantics. Samples from joint dataset from four main human body regions are displayed in the form of 2D slices in Figure 2 **c**. To enhance the spatial segmentation capabilities of SegVol, FH algorithm[30] is performed to generate 510k pseudo volumetric mask labels to fill in unannotated regions for these instances. Additionally, in order to build a universal feature extractor for volumetric medical images, we collect 90k unlabeled open-source CTs for pre-training. These data and annotations form the basis of SegVol.

The top 30 major categories of the joint dataset is present in Figure 2 **b**. For anatomical structures with multiple substructures such as rib, vertebrae, and lung, we treat them as a single unit for counting, thus giving them a significant advantage in mask quantity. Apart from these composite targets, the predominant categories in the mask labels of the joint dataset are important organs such as kidney, heart, liver, pancreas, adrenal, and spleen. Additionally, Figure 2 **b** summarizes the distribution of mask labels in the four main parts of the human body within the joint dataset. The Thorax and Abdomen contain relatively abundant anatomical structures, accounting for 39.2% and 44.9% mask labels of the entire joint dataset respectively. The Head and Neck and Pelvis sections are relatively less emphasized in medical image segmentation field, resulting in a smaller proportion of mask labels in the collected joint dataset, accounting for only 3.6% and 12.2%, respectively. Additionally, there are some categories cannot be classified into a specific body part, such as skin, which are not displayed in the Figure 2 **b** and represent a very small proportion of mask labels.

During the training and testing process, each subset of the joint dataset is divided into 80% training data and 20% testing data. In order to ensure the absence of any

data leaks, hash value is utilized to compare between the test set and the training set. We use the Dice Similarity Coefficient (Dice score) as metric to evaluate model, which is defined as $DSC = \frac{2|X \cap Y|}{|X|+|Y|}$. $|X \cap Y|$ is the cardinality of the intersection of the predicted segmentation sets X and the ground truth sets Y. $|X|$ and $|Y|$ are the cardinalities of sets X and Y respectively. Dice score is a commonly used metric for evaluating image segmentation tasks. It measures the degree of similarity between predicted segmentation and true segmentation, making it particularly suitable for evaluating the overlap degree of binary segmentation results.

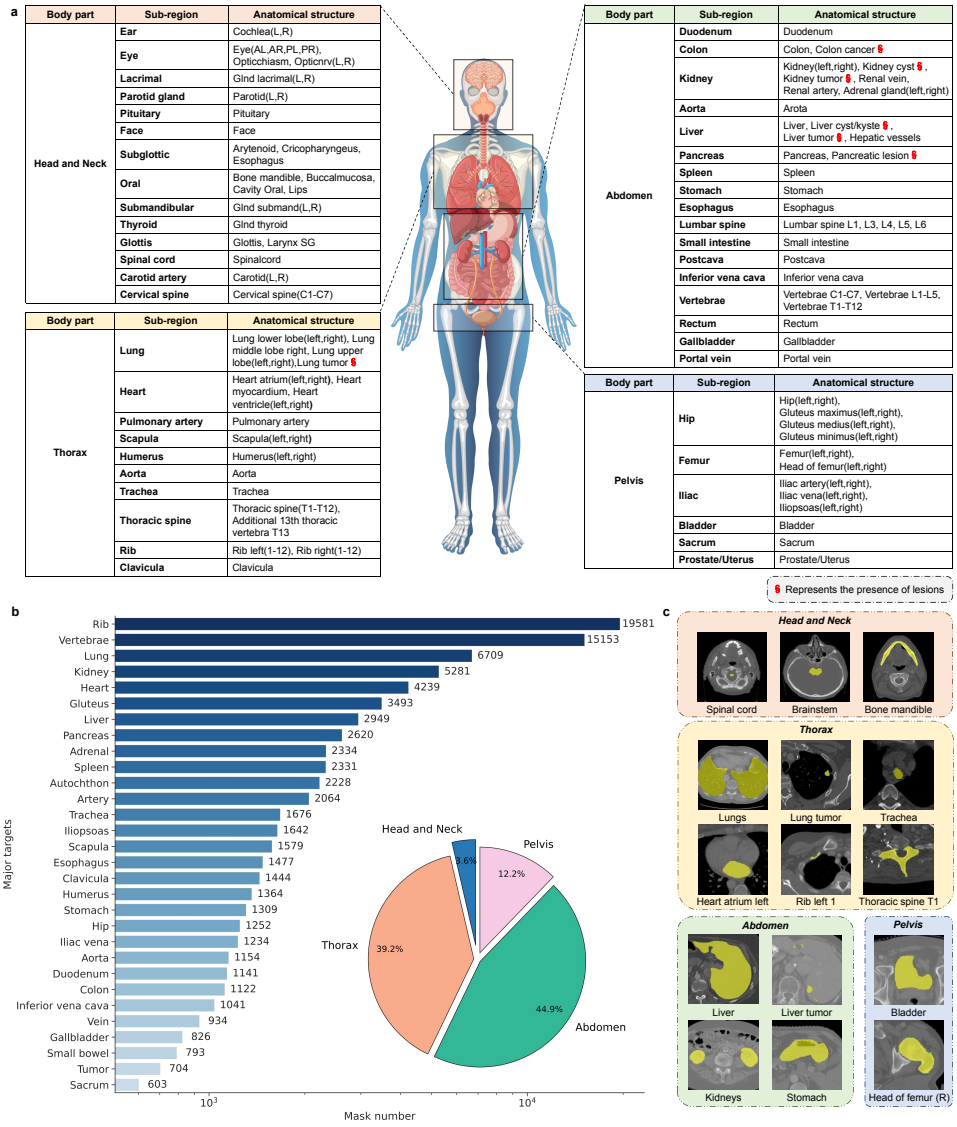

Fig. 2 An overview and examples of the joint dataset. **a** Overview of the joint dataset. The joint dataset comprises 47 important regions, with each region containing one or multiple significant anatomical structures within that spatial area. **b** Main categories of the joint dataset: their mask label quantities rank in the top 30 and proportion of mask label counts in the four main parts of the human body within the joint dataset. **c** Examples of organs, tissues and lesions from 15 different categories sampled from the joint dataset, presented in a slice view. Image of human body by brgfx on Freepik[31].

Table 2 Dice score(%) comparative analysis with traditional volumetric medical image segmentation methods.

	BTCV[23]		**Lung***	**Spleen***	**Colon***	**Liver***		**Average**	
Methods	Easy	Hard	Hard	Easy	Hard	Easy	Hard	Easy	Hard
nnFormer[6]	64.88	12.47	38.62	74.08	18.86	88.68	47.24	75.88	29.30
SwinUNETR[7]	74.32	25.72	44.06	54.54	9.02	91.75	51.52	73.54	32.58
3DUX-Net[10]	74.46	34.06	38.32	60.20	12.39	91.74	50.88	75.47	33.91
nnU-Net v2[9]	89.71	77.06	59.63	97.27	37.69	87.16	63.06	91.38	59.36
OURS	93.87	77.37	71.22	96.21	69.65	94.87	78.25	94.98	74.12

Rows headers marked with * represent the subset of the MSD dataset[21].

For detailed information on category splits of easy/hard, please refer to Table 5.

2.2 Segmentation Results of Major Targets

Under the support of prompt learning, SegVol is able to segment over 200 categories. We select 19 important anatomical targets to demonstrate its powerful segmentation capabilities, as illustrated in Table 1. SegVol demonstrates its remarkable ability to segment these vital targets. It achieves a remarkable Dice score of up to 96.13% for liver, while registers impressive 83.02% for the 19 major targets on average. Its powerful universal segmentation capabilities come from composite-type prompts with both semantic and spatial perspectives. On the one hand, spatial prompts allow the model to comprehend the specific space and location referred to. According to Table 1, the Dice score of 'box+text' prompt is boosted 5.85% compared to the text prompt on the average level of various organ segmentation results. On the other hand, semantic prompts clarify the reference to the anatomical structure, eliminating multiple plausible outcomes. This is reflected in Table 1 as the average Dice score of 'point+text' prompts is 4.62% higher than using point prompts alone. Spatial and semantic prompts mutually support each other, ultimately endowing the model with powerful segmentation capabilities.

2.3 Performance Comparison with Traditional Baseline Models

Segmentation models mainly fall into two architectures, CNN-based models and Transformer-based models. We conduct comparative experiments with representative CNN-based models such as 3DUX-Net(ICLR)[10] and nnU-Net(Nature methods)[9], and representative Transformer-based models such as nnFormer[6] and SwinUNETR[7]. In order to evaluate the models ability of organs, tissues, and lesions segmentation, we perform comparative experiments on BTCV[23] and MSD-spleen[21] datasets, which focus on organ segmentation problems, and on MSD-lung, MSD-colon and MSD-liver datasets, which focus on lesion segmentation problems. Due to the significant differences in the model's segmentation response for 'MegaStructures' and 'MicroStructures' anatomical targets, we split the categories into easy groups and hard groups to better understand the performance differences of different models. For detailed information on category splits, please refer to Table 5. In the comparative

Table 3 Dice score(%) results for lesion segmentation.

Methods	**MSD-lung**[21] Lung tumor	**MSD-colon** Colon cancer	**MSD-liver** Liver tumor	**Average**
nnU-Net v2[9]	59.63	37.69	63.06	53.46
OURS	71.22	69.65	78.25	73.04

experiments, the original open-source training set is devised into 80% for training and 20% for testing. The experimental results are summarized in Table 2.

For classic volumetric medical image datasets ranging from dozens to hundreds of cases, SegVol jointly trained on 25 datasets, significantly outperforms traditional segmentation models trained on a single dataset. Table 2 illustrates that SegVol surpasses traditional models in easy categories such as liver, kidney, spleen, etc., with a Dice score reaching 94.98%. This is mainly due to its more knowledge learned from the same categories of other datasets. More significantly, our method maintains a leading position in segmentation of hard categories such as liver tumor, lung tumor, and adrenal gland. The average Dice score of SegVol for hard categories is 14.76% higher in absolute term than that of nnU-net[9], which ranks second, performing precise segmentation of hard samples. The reason is SegVol can obtain prior information through spacial and semantic prompts, thereby enhancing the understanding of hard samples and significantly improving segmentation results.

We analyze that there are mainly three factors that make SegVol significantly outperform traditional models: 1) Finite cases of fix class set limit the performance of traditional models. Conversely, SegVol has a broader learning scope as it combines 25 datasets for training. This allows it to glean knowledge not only from identical categories across different datasets but also from categories that are inherently correlated within the embedding space of natural language. For instance, SegVol can learn from both 'left kidney' and 'kidney' categories due to their natural language correlation. This ability to learn from a wider and more diverse range of data makes an edge over traditional models, enabling it to understand the intrinsic correlations of segmentation targets that traditional models might miss. 2) While traditional models find out semantic information solely relying on integer codes, SegVol takes a more comprehensive prompt learning approach. It not only leverages semantic prompts to understand targets but also utilizes spatial prompts to gain further spatial cognition of the targets. This interactive segmentation mode enables SegVol to achieve significantly better results, particularly in the precise segmentation of hard cases. 3) Pre-training on large-scale unlabeled data makes SegVol capable of learning more generalized feature representations. This process significantly enhances its adaptability and robustness in downstream tasks.

2.4 Evaluating Lesion Segmentation Capability

Segmenting lesion volumetric structures precisely is of significant importance in clinical medical applications. Three key datasets, MSD-lung, MSD-colon, and MSD-liver[21], are selected to evaluate the ability of SegVol in segmenting lesion anatomical structures. The MSD-lung dataset comprises patients diagnosed with non-small cell lung

Table 4 Dice score(%) results of ablation study on Zoom-out-zoom-in mechanism applied to MSD-liver dataset.

		Single-type prompt			Composite-type prompt		
Mechanism	**Category**	Text	Point	Box	Point+Text	Box+Text	**Average**
Resize	Liver	87.11	87.29	87.60	86.55	87.52	87.21
Zoom-out-zoom-in	Liver	95.07	87.92	94.15	94.37	94.87	93.28
Resize	Liver tumor	45.48	51.82	50.67	53.54	55.24	51.35
Zoom-out-zoom-in	Liver tumor	75.73	72.51	62.37	74.51	78.25	72.67

Table 5 Dataset categories divided into easy and hard classes.

Dataset	**Easy categories**	**Hard categories**
BTCV[23]	Spleen, Liver, Stomach, Right kidney, Left kidney,	Gallbladder, Esophagus, Aorta, Inferior vena cava, Portal vein and splenic vein, Pancreas, Right adrenal gland, Left adrenal gland
MSD-lung[21]	-	Lung tumor
MSD-spleen	Spleen	-
MSD-colon	-	Colon cancer
MSD-liver	Liver	Liver tumor

cancer from Stanford University (Palo Alto, CA, USA). The MSD-colon dataset includes patients who underwent resection of primary colon cancer at the Memorial Sloan Kettering Cancer Center (New York, NY, USA). The MSD-liver dataset encompasses a variety of primary cancers, including hepatocellular carcinoma, as well as metastatic liver disease derived from colorectal, breast, and lung primary cancers. The CT scans in this dataset include a range of pre-therapy and post-therapy images[21].

we use the nnU-net[9] as baseline model, which show strongest segmentation ability in traditional volumetric medical image segmentation models. As demonstrated in Table 3, the ability of SegVol to segment these challenging lesion cases is markedly superior to that of nnU-net. Across these three lesion datasets, the Dice score of SegVol exceeds that of nnU-net by 19.58% in absolute term, representing a significant advancement in the segmentation of complex volumetric cases. Figure 5 **c** presents a series of examples illustrating the lesion segmentation performance of both nnU-net and our method. These examples encompass liver tumors, colon cancer, and lung tumors. The visualization results reveal that these lesion anatomical structures reconstructed by SegVol align more closely with the ground truth compared to the results produced by nnU-net.

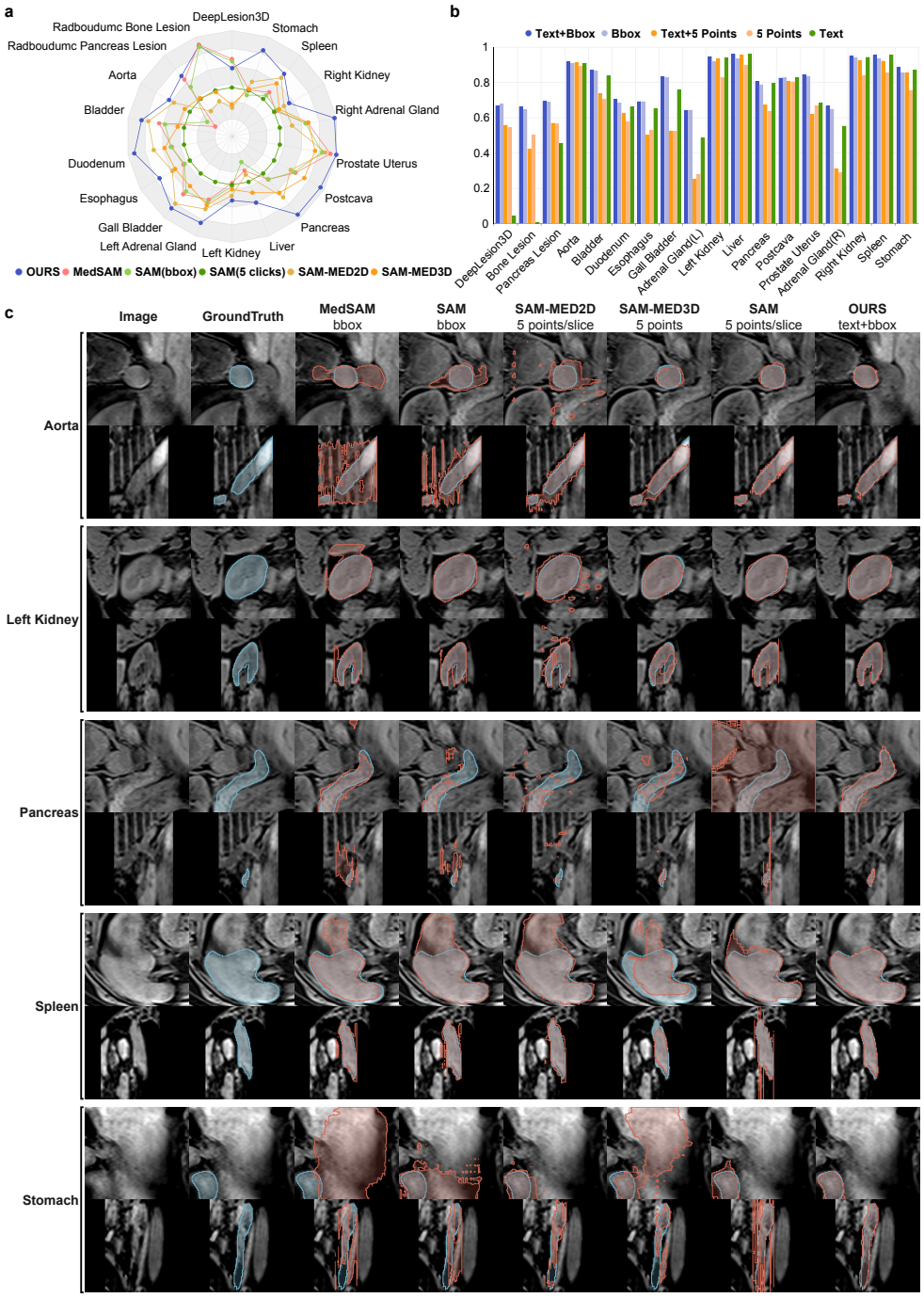

Fig. 3 Evaluation results on the external validation sets for interactive models. **a** Radar map which shows the external validation results of interactive models. Each axis represents a category in external set and each data point represents the relative dice average score compared to standard baseline, SAM(5 clicks). **b** Performance consistency of SegVol in external validation experiment among different prompt types. Clusters of the bar chart correspond to the categories in external set. Different bars in clusters represent different prompt types for SegVol. **c** Visualized segmentation results from 5 interactive models on the external validation set. In each case, the upper row is axial plane of CTs and the lower row is sagittal plane of CTs.

2.5 External Validation

To compared with other interactive segmentation models [32–35], we performed external validation experiments on 1,738 cases from the validation set of AMOS22[36] and the whole novel annotated set of Universal Lesion Segmentation Challenge 23(ULS23)[37]. The validation set of AMOS22 contains 120 cases annotated with 15 important organs. The novel annotated ULS23 dataset is composed with DeepLesion3D subset, Radboudumc Bone subset and Radboudumc Pancreas subset. The DeepLesion3D subset contains 200 abdominal lesions, 100 bone lesions, 50 kidney lesions, 50 liver lesions, 100 lung lesions, 100 mediastinal lesions and 150 assorted lesions cases. There are 744 bone lesion cases in Radboudumc Bone subset and 124 pancreas lesion cases in Radboudumc Pancreas subset. During the external validation process, the models' parameters are frozen and dice score is used to evaluate the generalization ability of these interactive models.

The results of external validation experiments in showed in Figure 3. The Figure 3 **a** illustrates our methods is in the leading position in most of domains including lesions and organs compared with other SAM-like interactive models. MedSAM[33] and SAM(bbox)[32] is applied bounding box prompts. SAM(5 clicks)[32], SAM-MED2D[34] and SAM-MED3D[35] is applied point prompts using step correction for 5 times. Step correction means that point prompt in each step will be given according to the previous output and ground truth, rather than giving it all at once. Our method is applied bounding box and text prompt. The Figure 3 **b** shows the performance consistency of SegVol among different prompt types, such as bbox prompt and text+bbox prompt. Noticed that category of each mask in ULS23 is not clearly defined. So we have to give a general text(like tumor or lesion) to prompt SegVol and it still demonstrates strong competitiveness compared to other interactive models.

In Figure 3 **c**, we visualize the segmentation results in 5 important organ categories to study the difference within these interactive models. Due to the lack of understanding of 3D space, the 2D methods like MedSAM and SAM(bbox) are unable to capture the complex 3D spatial information of targets, resulting in relative worse results. The 3D methods, SAM-MED3D, performs well in aorta case and stomach case, but demonstrates poor segmentation ability in pancreas case. In contrast, SegVol is stable in all these categories, relying on its understanding of 3D structures and semantic clarification of text prompt.

2.6 Case Study

2.6.1 Eliminate Multi Plausible Outputs

Alexander Kirillov, etc.[32] discuss the multiple plausible outputs problem in spatial prompt setting. Such as three sub-figures on top left in Figure 4 **a**, they are applied the same point prompt on the kidney structure. But kidney tumor, left kidney and both left kidney and right kidney are all reasonable outputs. Or such as bottom left, the bounding box selects the region of liver. However, liver tumor, hepatic vessels and liver itself are also plausible target structures. In these cases, SAM chooses to return multiple masks to match different levels of plausible results. Unlike SAM's solution, we use semantic prompts to clarify the targets. As shown in Figure 4 **a**, the captions

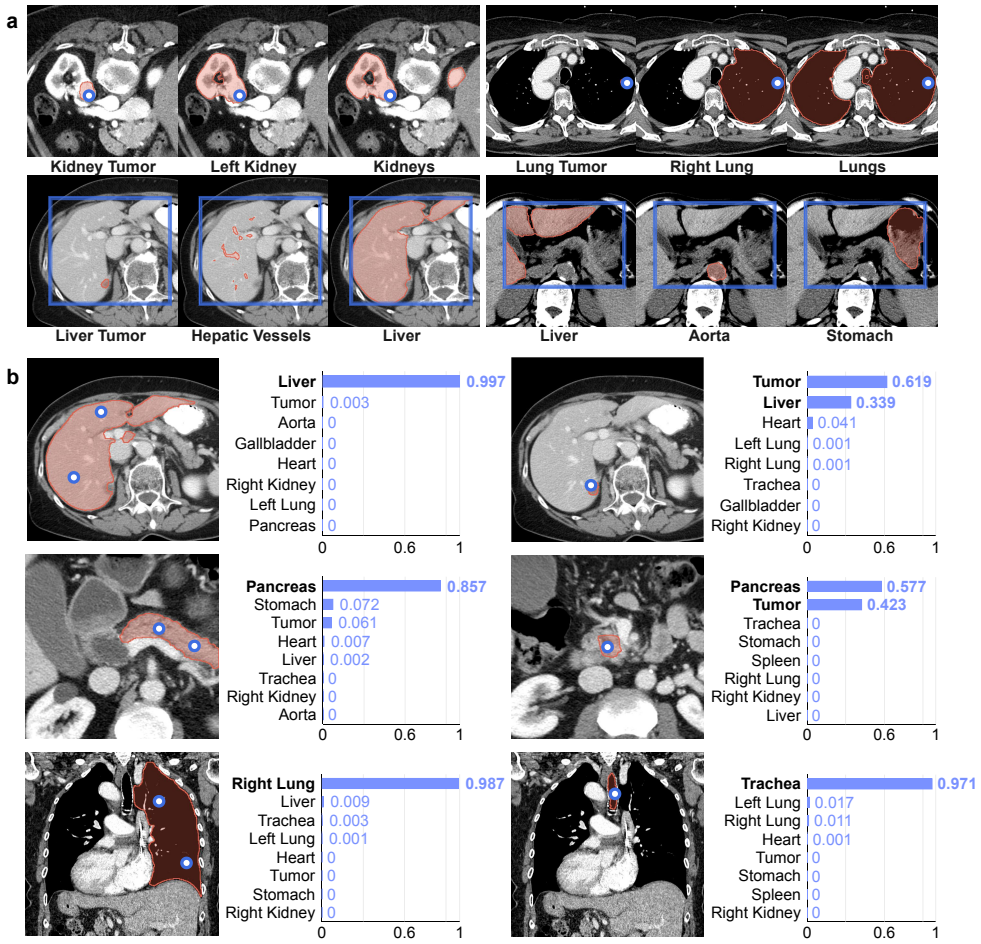

Fig. 4 Case studies on semantic clarification ability and spatial-semantic reflection ability of SegVol. **a** Four cases for semantic prompts clarify semantic ambiguity to eliminate multi plausible outputs. The image rows show the spatial prompts and mask predictions generated by SegVol. The caption rows show the semantic prompts corresponding to each image. **b** Six cases for reflection ability of SegVol to transfer spatial prompt to semantic categories. The image lines show the spatial prompts and mask predictions. The bar chart lines show the top 8 semantic categories and their predicted scores.

below the images are text prompts of model inputs, and the masks in the images is the predictions of SegVol. The visualization results show SegVol based on the same spatial prompt can generate accurate corresponding predictions relying on different semantic prompts.

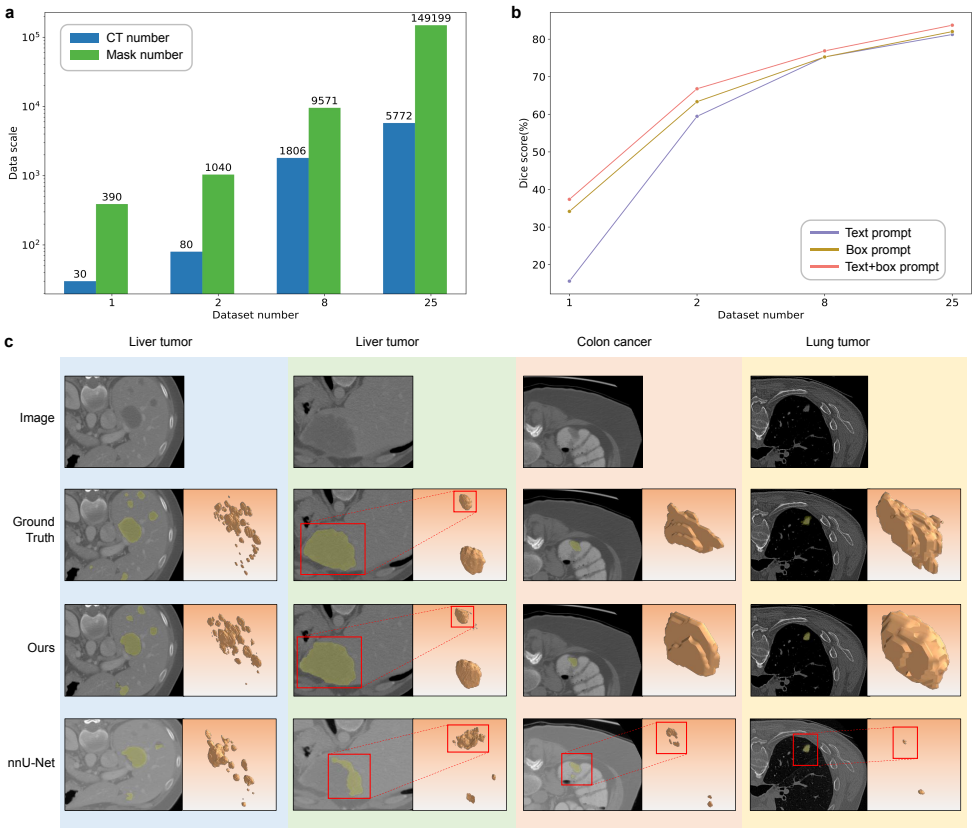

Fig. 5 Experimental results analysis for data quantity scale and lesion segmentation. **a** Bar graph representation of Computed Tomography (CT) quantities and corresponding ground truth mask quantities in various number of datasets. **b** Line graph depicting Dice scores of models trained on various datasets. **c** Visual representation of lesion segmentation results.

2.6.2 Transfer Spatial Prompt to Semantic Categories

Furthermore, we study the possibility of SegVol to reflect spatial prompt to semantic categories. Figure 4 **b** demonstrates that SegVol can give accurate semantic categories based on the spatial prompts. The top left sub-figure in Figure 4 **b** with the spatial prompt on liver make SegVol get 0.997 prediction score for liver. And the top right sub-figure shows if the spatial prompt is point on the liver tumor, SegVol will get 0.619 prediction score for tumor category and 0.339 prediction score for liver because of the spatial relationship of liver tumor and liver. We implement this reflection by reusing the SegVol decoder with semantic prompts from a category set to get logits of different categories in the region of interesting and apply softmax function to get the prediction ranking.

2.7 Ablation Study

2.7.1 Zoom-out-zoom-in Mechanism

The computation of volumetric images segmentation is a resource-consuming process. Traditional methods for volumetric medical image segmentation often employ a sliding window approach to mitigate computational expenses. However, this approach is time-consuming and confines the model to infer only from localized parts of image, thereby losing critical global structural information. This shortcoming is particularly significant for large anatomical structures such as the liver or kidney. We introduce a Zoom-out-zoom-in mechanism to address it, which is illustrated in Figure 6 **c** and **d**. This innovative method cuts the computational cost while simultaneously supporting full-size input and inference at original resolution.

An ablation study is conducted on MSD-liver dataset[21] to evaluate the contribution of the Zoom-out-zoom-in mechanism. The MSD-liver dataset comprises categories for both liver and liver tumor, allowing to investigate the impact of the Zoom-out-zoom-in mechanism on both 'MegaStructures' and 'MicroStructures' targets. As detailed in Table 4, the application of the Zoom-out-zoom-in mechanism to the SegVol model results in a 6.07% Dice score increase for the liver category. The improvement is even more pronounced in the liver tumor category, where the Zoom-out-zoom-in mechanism boost the Dice score of SegVol by 21.32%. Interestingly, the application of Zoom-out-zoom-in mechanism result in only a slight improvement within the point prompt setting for liver target. We propose it could be attributed to the relative scarcity of point prompts at a global level, which becomes more noticeable when the focus is narrowed to a local region, thereby limiting the potential for increase.

2.7.2 Dataset Scale

One of crucial factor n the construction of foundation models is the scale of data. We conduct an ablation study to investigate the impact of the quantity of images and masks on the performance of the model. The BTCV dataset[23], which includes 13 significant organs, is set as anchor to evaluate the model trained separately on 1, 2 and 8 datasets for 500 epochs, as well as the final model trained on 25 datasets. The detailed results are shown in Figure 5 **a** and **b**. As a lightweight model, the performance is suboptimal when only one dataset is used. However, with the increase in data quantity, the Dice score of the model increases significantly, especially in the text prompt setting, which heavily relies on the number of ground truth masks with semantic information.

3 Discussion

We present SegVol, a foundational model for interactive, universal volumetric medical image segmentation. This model has been developed and evaluated using 25 open-source datasets. Unlike the strongest traditional volumetric segmentation method, nnU-net[9], which automatically configures settings for each dataset, SegVol is designed to unify various volumetric segmentation datasets into a single architecture. This results in a universal segmentation tool capable of generating accurate responses

for over 200 anatomical targets. Furthermore, SegVol demonstrates state-of-the-art or near state-of-the-art volumetric segmentation performance when compared with traditional methods[5–10], particularly for lesion anatomical targets. Despite its universality and precision, SegVol maintains a lightweight architecture compared to other volumetric segmentation methods. We have made SegVol an open-source foundational model, readily applicable to a broad spectrum of medical image representation and analysis fields. This ensures it can be easily integrated and utilized by researchers and practitioners alike.

Traditional volumetric image segmentation methods[5–10], due to their formulation that relies on integer codes to represent semantic information, are incapable of modeling the interrelation between various anatomical categories. To address this, we use prompt learning technology to drive SegVol. Semantic prompts are employed to model various anatomical categories in a natural language embedding space and spatial prompts are used to specifically refer to the spacial structure of organs, tissues, and lesions. The ability to accept prompts makes SegVol a universal and precise model for volumetric medical image segmentation. To mitigate the computational cost associated with volumetric images, we propose a Zoom-out-zoom-in mechanism. It allows users to simply prompt the model in global view, while receiving a quick response in the original resolution. This innovative approach ensures both efficiency and precision in medical image analysis.

While SegVol is capable of understanding semantic prompts composed of sentences, there remains a gap between it and the referring expression segmentation that involves complex semantic information and logical relationships. The establishment of a referring expression segmentation model needs more related data with logical relationships, rather than merely segmentation datasets. Furthermore, as a closed-set model, SegVol has a limited ability to segment unseen categories. However, we remain optimistic about this limitation. The current framework allows for the direct addition of new data in the same format, even if the new data includes categories that have never been seen before. This means that the model can inherit all previous knowledge and continue learning in new fields. This adaptability and continuous learning capability make SegVol a promising tool in the field of medical image segmentation.

We primarily use SegVol with Computed Tomography (CT) data due to its advantages such as fast image acquisition, clear details, and high contrast resolution. CT is also the preferred method for evaluating solid tumors. However, the flexible architecture of SegVol allows it to be compatible with various types of volumetric medical images, like MRI. We believe this versatility could enable SegVol to be applied broadly in the representation and analysis of volumetric medical images.

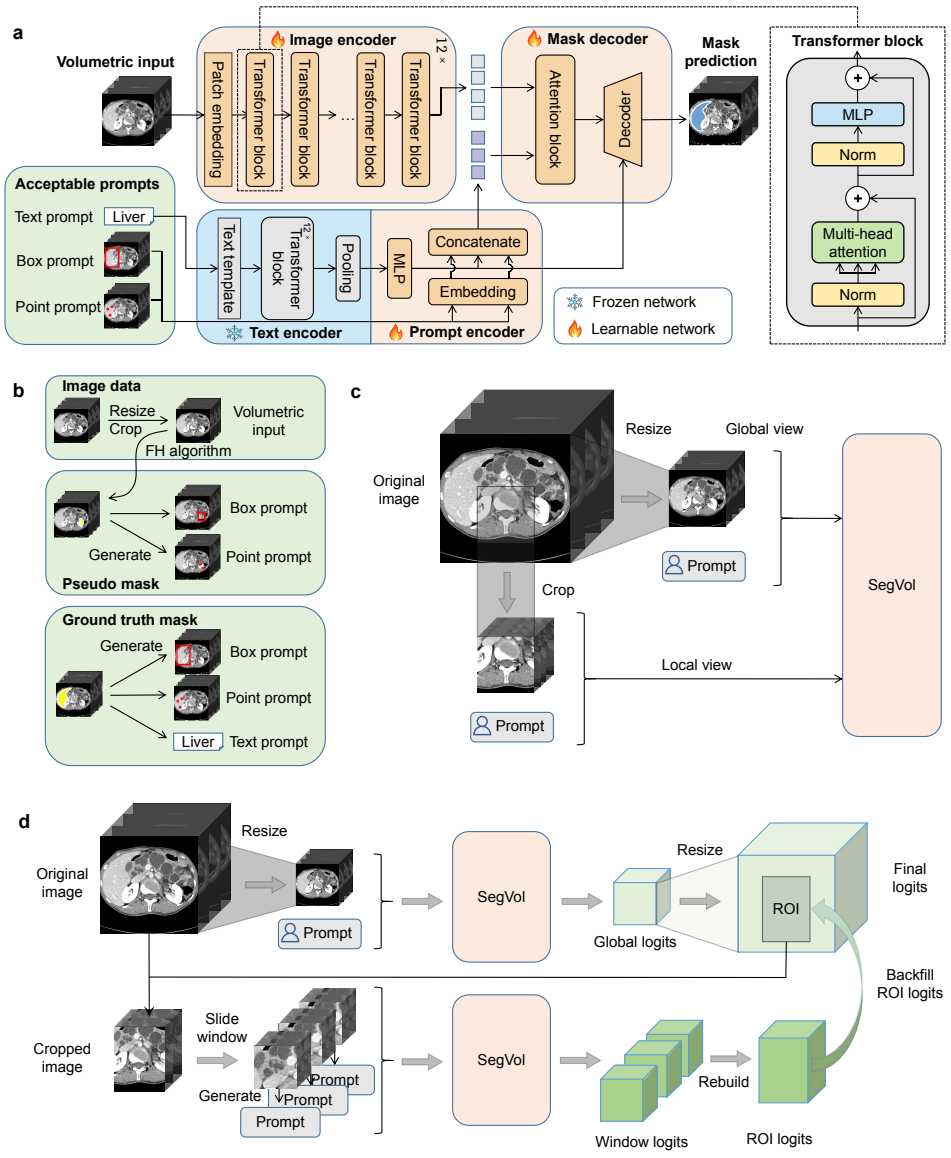

Fig. 6 An overview of the model architecture of the proposed SegVol. **a** The main structure of SegVol includes the image encoder, text encoder, prompt encoder, and mask decoder. All networks, except the text encoder, are learnable. The image encoder extracts the image embedding of volumetric input. The image embedding is fed into the decoder together with prompt embeddings to predict the segmentation mask. **b** Illustration of input image transformation and the prompt generation. **c** Zoom-out-zoom-in training: SegVol is trained on data of both global and local views. **d** Zoom-out-zoom-in inference: SegVol first conducts global inference and then performs local inference on the extracted ROI to refine the results.

4 Method

4.1 Data Processing

4.1.1 Data Collection and Normalization

One of main challenges in volumetric medical image segmentation is the absence of large-scale publicly available volumetric medical data, especially the annotated segmentation CTs. Doing our utmost, we collected 25 open-source segmentation CT datasets, including CHAOS[38–40], HaN-Seg[17], AMOS22[36], AbdomenCT-1k[41], KiTS23[42], KiPA22[26–29], KiTS19[43], BTCV[23], Pancreas-CT[14, 44, 45], 3D-IRCADb[46], AbdomenCT-12organ[21, 22], TotalSegmentator[24], CT-ORG[11–14], VerSe19, VerSe20[18–20], SLIVER07[47], QUBIQ[48], six MSD datasets[21], LUNA16[15], and WORD[25]. These CTs originates from various medical institutions, captured by different machines with varying parameter settings and scanning regions. These factors result in remarkable differences in data distribution, and thus significant challenge in data processing.

To standardize these datasets, we perform the following transformation on every CT scan: We first compute a threshold based on the mean voxel value of each volume. Voxels with values that are above this threshold are considered as the foreground. We calculate the 99.95^{th} and 0.05^{th} percentiles for the foreground voxels, and use them as the upper and lower bounds to clip the original voxels. We further normalize the foreground voxels using the mean and standard deviation.

4.1.2 Pseudo Mask Generation and De-noising

Volumetric segmentation datasets suffer from the notorious problem of partial label. Most of the datasets have annotations of only a few segmentation targets, e.g., several organs. Therefore, the deep models may learn the spurious correlation between datasets and segmentation targets, and produce inferior results during inference phase. To eliminate this problem, we utilize the Felzenswalb-Huttenlocher (FH)[30] algorithm to generate pseudo masks for most of the objects in each CT scan.

The unsupervised segmentation algorithm FH[30] separates the spatial structures based on the gradient between adjacent voxels. However, pseudo masks derived by FH algorithm contain substantial noise and numerous small masks. The algorithm may also cause inaccurate segmentation, such as the disconnection of some continuous structure and the wrong connection of different structures. To improve the pseudo masks, we employ the following strategies: 1) The pseudo masks are replaced with ground truth masks when applicable. 2) We filter out tiny structures smaller than 1‰ of the whole volume. 3) Each mask is refined by dilation and erosion operations.

4.2 Model Architecture

The volumetric medical image segmentation dataset $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$ consists of a 3D image datum $\boldsymbol{x}_i \in \mathcal{R}^{C\times D\times H\times W}$ and K mask labels $\boldsymbol{y}_i \in \{0,1\}^{K\times D\times H\times W}$, corresponding to K target categories. The classic segmentation model[5–10] $\mathcal{F}(*, \boldsymbol{\theta})$ learns to predict masks $\boldsymbol{y}_i$ belonging to the K categories based on the volumetric input $\boldsymbol{x}_i$,

i.e., $\boldsymbol{o}_i = \mathcal{F}(\boldsymbol{x}_i, \boldsymbol{\theta})$, where $\boldsymbol{o}_i \in \mathcal{R}^{K\times D\times H\times W}$. Therefore, the traditional models are not able to generalize to unseen categories.

Motivated by the recent advance in 2D nature image segmentation, Segment Anything (SAM)[32], we design a novel method for interactive and universal volumetric medical image segmentation, named, SegVol. We illustrate the model in Fig. 6 **a**. SegVol supports three types of prompts for interactive segmentation: 'box' prompt, $\boldsymbol{b} \in \mathcal{R}^6$ representing the coordinates of two diagonal vertices; 'point' prompt, including a set of (P) points $\boldsymbol{p} \in \mathcal{R}^{P\times 3}$; and 'text' prompt, such as 'liver' or 'cervical spine C2', which is tokenized to tensor $\boldsymbol{t}$. SegVol consists of four modules, namely, image encoder $\mathcal{F}_{\text{IE}}(*, \boldsymbol{\theta}_{\text{IE}})$, text encoder $\mathcal{F}_{\text{TE}}(*, \boldsymbol{\theta}_{\text{TE}})$, prompt encoder $\mathcal{F}_{\text{PE}}(*, \boldsymbol{\theta}_{\text{PE}})$, and mask decoder $\mathcal{F}_{\text{MD}}(*, \boldsymbol{\theta}_{\text{MD}})$. We introduce each module in the following.

Image encoder. We employ ViT (Vision Transformer)[49] as the image encoder, which exhibits remarkable advantages over convolutional models[50] when pre-trained on large-scale datasets. We first pre-train ViT using MAE algorithm[51] on the all collected 96k CTs, and then conduct further supervised training on the 6k CTs with 150k labeled segmentation masks. The image encoder, denoted as $\mathcal{F}_{\text{IE}}(*, \boldsymbol{\theta}_{\text{IE}})$, takes a volumetric image $\boldsymbol{x} \in \mathcal{R}^{C\times D\times H\times W}$ as input. Firstly, it splits $\boldsymbol{x}$ into a set of patches, denoted as $\boldsymbol{x}_{\text{patch}} \in \mathcal{R}^{N\times(C\times P_D\times P_H\times P_W)}$, where $N = \frac{D\times H\times W}{P_D\times P_H\times P_W}$. P_D, P_H and P_W are the size of patch. These patches are then fed into the network, which outputs an embedding $\boldsymbol{z}_{\text{image}} = \mathcal{F}_{\text{IE}}(\boldsymbol{x}_{\text{patch}}, \boldsymbol{\theta}_{\text{IE}})$, $\boldsymbol{z}_{\text{image}} \in \mathcal{R}^{N\times F}$. F represents the feature dimension, which is set to 768 by default in this paper.

Text prompt encoder. One main limitation of traditional segmentation models is that the models learn dataset-specific labels encoded as integers which cannot generalized to new datasets or tasks, limiting its real-world applications. We enable universal segmentation cross datasets by leveraging the text prompt. We employ the text encoder from CLIP model[52] to encode the input text prompt, as CLIP[52] has been trained to align image and text on web-scale image-text pairs. We denote the text prompt encoder as $\mathcal{F}_{\text{TE}}(*, \boldsymbol{\theta}_{\text{TE}})$. Given a word or or phrase as prompt, we complete it using the template $\boldsymbol{s}$ ='A computerized tomography of a [text prompt]'[53]. $\boldsymbol{s}$ is then tokenized into $\boldsymbol{t}$. The text encoder accepts $\boldsymbol{t}$ as input and outputs the text embedding $\boldsymbol{z}_{\text{text}} = \mathcal{F}_{\text{TE}}(\boldsymbol{t}, \boldsymbol{\theta}_{\text{TE}})$, where $\boldsymbol{z}_{\text{text}} \in \mathcal{R}^F$. We freeze the off-the-shelf text encoder during training, since the text data in CT datasets is of a small amount.

Spatial prompt encoder. Following SAM[32], we use the positional encoding[54] for point prompt $\boldsymbol{p}$ and box prompt $\boldsymbol{b}$ and obtain the point embedding $\boldsymbol{z}_{\text{point}} \in \mathcal{R}^F$ and box embedding $\boldsymbol{z}_{\text{box}} \in \mathcal{R}^F$. We concatenate the embeddings of three kinds of prompts as $\boldsymbol{z}_{\text{prompt}} = \mathcal{F}_{\text{PE}}(\boldsymbol{p}, \boldsymbol{b}, \boldsymbol{s}, \boldsymbol{\theta}_{\text{PE}}) = [\boldsymbol{z}_{\text{point}}, \boldsymbol{z}_{\text{box}}, \boldsymbol{z}_{\text{text}}]$.

Mask decoder. After obtaining the image embedding $\boldsymbol{z}_{\text{image}}$, prompt embedding $\boldsymbol{z}_{\text{prompt}}$ and text embedding $\boldsymbol{z}_{\text{text}}$, we input them to the mask decoder and predict the mask $\boldsymbol{p} = \mathcal{F}_{\text{MD}}(\boldsymbol{z}_{\text{image}}, \boldsymbol{z}_{\text{prompt}}, \boldsymbol{z}_{\text{text}}, \boldsymbol{\theta}_{\text{MD}})$. We use self-attention and cross-attention[55] in two directions to blend the image embedding and prompt embedding, and then employ the transposed convolutions and interpolation operations to generate masks. Since the text embedding is the key to universal segmentation and it is

also harder to learn the correlation between text and volumetric regions, we reinforce the text information by introducing a parallel text input $\boldsymbol{z}_{\text{text}}$ beside the joint prompt embedding $\boldsymbol{z}_{\text{prompt}}$. We further compute a similarity matrix between the up-scaled embedding from the transposed convolution output and the text embedding in the mask decoder. The element-wise multiplication of the similarity matrix with the mask prediction is applied before interpolation, after which the model outputs the masks.

4.3 Training procedure

Prompt generation. SegVol can accept multiple prompt types, including individual point prompts, box prompts and text prompts, and also their combinations. To make full use of the segmentation training data, we generate kinds of prompts for each datum. Then, the prompt and mask pairs are used to compute the training loss. SegVol supports 'point' prompts, 'box' prompts and 'text' prompts.

The point prompt is built from ground truth or pseudo masks, consisting of three kinds of points, namely, positive point, negative point, and ignore point. Positive point means that it is within the target mask region, while negative points are those outside. The ignore points are used for the purpose of input completion, which will be disregarded by the model, so that the point prompt has the same length.

The box prompt is generated based on the ground truth or pseudo masks, integrated with random jitter to enhance the model's robustness. When generating the box prompt for some pseudo mask, the box may also cover other masks due to the irregular 3D shapes. To address this problem, we compute the Intersection over Union (IOU) between the generated box and the included pseudo masks. Any mask with an IOU greater than 0.9 will also be integrated and considered as part of the target mask corresponding to this box prompt.

The box and point prompts can be generated by sampling points based on the ground-truth segmentation masks, while text prompts are constructed based on their category names. As pseudo masks produced by the unsupervised FH algorithm[30] do not have the semantic information, we only use point and box prompts (Fig. 6 **b**) for training with pseudo masks.

Loss function. We combine the binary cross-entropy (BCE) loss and Dice loss as the loss function $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$ to train the model with parameters $\boldsymbol{\theta}$, where $\boldsymbol{\theta} = [\boldsymbol{\theta}_{\text{IE}}, \boldsymbol{\theta}_{\text{PE}}, \boldsymbol{\theta}_{\text{MD}}]$ and $\mathcal{D}$ is the training dataset. The loss function is as follows:

$$\mathcal{L}_{\text{BCE}}(\boldsymbol{\theta}; \mathcal{D}) = -\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}[\langle \boldsymbol{y}, \log(\mathcal{F}(\boldsymbol{x}, \boldsymbol{\theta}))\rangle + \langle 1 - \boldsymbol{y}, \log(1 - \mathcal{F}(\boldsymbol{x}, \boldsymbol{\theta}))\rangle] \tag{1}$$

$$\mathcal{L}_{\text{Dice}}(\boldsymbol{\theta}; \mathcal{D}) = 1 - \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}[\frac{2\langle \boldsymbol{y}, \mathcal{F}(\boldsymbol{x}, \boldsymbol{\theta})\rangle}{\|\boldsymbol{y}\|_1 + \|\mathcal{F}(\boldsymbol{x}, \boldsymbol{\theta})\|_1}] \tag{2}$$

$$\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) = \mathcal{L}_{\text{BCE}}(\boldsymbol{\theta}; \mathcal{D}) + \mathcal{L}_{\text{Dice}}(\boldsymbol{\theta}; \mathcal{D}) \tag{3}$$

4.4 Zoom-out-zoom-in Mechanism

Comparing to 2D slides, volumetric data has remarkably large number of voxels meanwhile small segmentation targets relatively. Naively down-sampling the original data will cause serious information loss and thus inferior performance. Diving the large volumetric data into small cubes and conquering each separately is computationally expensive and also suffers from information loss. To reduce the computational cost while preserving the details of the Region of Interest (ROI), we design the zoom-out-zoom-in mechanism consisting of multi-size training, zoom-out-zoom-in inference.

Multi-view training. To adapt various sizes of volumetric data and enable the zoom-out-zoom-in inference, we construct two kinds of training data. One is to resize the large-size CT to adapt the model's input size, and obtain the training data of zoom-out view. The other one is to crop the original large-size CT into cubes with model's input size. In this way, we obtain the training data of zoom-in view. The process is shown in Fig. 6 **c**.

Zoom-out-zoom-in Inference. Fig. 6 **d** illustrates our zoom-out-zoom-in inference. We first zoom-out and implement global inference. Given an large volumetric image, it is resized and then fed into SegVol model. After obtaining the global predicted segmentation mask based on user's prompt, we locate the region of interest (ROI) and zoom-in, namely, crop it from the original-size image. We apply sliding window on the cropped region and implement more precise local inference. We adapt the input prompt for the local inference, since the original point and box prompts input by the user may not be applicable in the local inference region when zoom-in. Specifically, we ignore positive or negative points out of the local region. Similar to the training box prompt generation in Sec. 4.3, we generate the local box prompt by considering the global predicted mask in the local region as the pseudo mask. Finally, we fill the ROI region of global segmentation mask with the local segmentation mask. The zoom-out-zoom-in mechanism realizes both efficient and precise inference simultaneously.

References

[1] Sajid, S., Hussain, S., Sarwar, A.: Brain tumor detection and segmentation in mr images using deep learning. Arabian Journal for Science and Engineering **44**, 9249–9261 (2019)

[2] Minnema, J., Ernst, A., Eijnatten, M., Pauwels, R., Forouzanfar, T., Batenburg, K.J., Wolff, J.: A review on the application of deep learning for ct reconstruction, bone segmentation and surgical planning in oral and maxillofacial surgery. Dentomaxillofacial Radiology **51**(7), 20210437 (2022)

[3] Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W., Rueckert, D.: Deep learning for cardiac image segmentation: a review. Frontiers in Cardiovascular Medicine **7**, 25 (2020)

[4] Samarasinghe, G., Jameson, M., Vinod, S., Field, M., Dowling, J., Sowmya, A., Holloway, L.: Deep learning for segmentation in radiation therapy planning: a review. Journal of Medical Imaging and Radiation Oncology **65**(5), 578–595 (2021)

[5] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 574–584 (2022)

[6] Zhou, H.-Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y.: nnFormer: Interleaved Transformer for Volumetric Segmentation (2022)

[7] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., Xu, D.: Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images (2022)

[8] 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 medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20730–20740 (2022)

[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] Lee, H.H., Bao, S., Huo, Y., Landman, B.A.: 3D UX-Net: A Large Kernel Volumetric ConvNet Modernizing Hierarchical Transformer for Medical Image Segmentation (2023)

[11] Rister, B., Shivakumar, K., Nobashi, T., Rubin, D.L.: Ct-org: Ct volumes with multiple organ segmentations [dataset]. The Cancer Imaging Archive (2019)

[12] Rister, B., Yi, D., Shivakumar, K., Nobashi, T., Rubin, D.L.: Ct organ segmentation using gpu data augmentation, unsupervised labels and iou loss. arXiv preprint arXiv:1811.11226 (2018)

[13] 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)

[14] 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)

[15] Setio, A.A.A., Traverso, A., De Bel, T., Berens, M.S., Van Den Bogaard, C.,

Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B., *et al.*: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. Medical image analysis **42**, 1–13 (2017)

[16] Jiang, H., Diao, Z., Yao, Y.-D.: Deep learning techniques for tumor segmentation: a review. The Journal of Supercomputing **78**(2), 1807–1851 (2022)

[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] Sekuboyina, A., Husseini, M.E., Bayat, A., Löffler, M., Liebl, H., Li, H., Tetteh, G., Kukačka, J., Payer, C., Štern, D., *et al.*: Verse: a vertebrae labelling and segmentation benchmark for multi-detector ct images. Medical image analysis **73**, 102166 (2021)

[19] Löffler, M.T., Sekuboyina, A., Jacob, A., Grau, A.-L., Scharr, A., El Husseini, M., Kallweit, M., Zimmer, C., Baum, T., Kirschke, J.S.: A vertebral segmentation dataset with fracture grading. Radiology: Artificial Intelligence **2**(4), 190138 (2020)

[20] Liebl, H., Schinz, D., Sekuboyina, A., Malagutti, L., Löffler, M.T., Bayat, A., El Husseini, M., Tetteh, G., Grau, K., Niederreiter, E., *et al.*: A computed tomography vertebral segmentation dataset with anatomical variations and multi-vendor scanner data. Scientific data **8**(1), 284 (2021)

[21] 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)

[22] Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., *et al.*: Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6695–6714 (2021)

[23] 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 VaultWorkshop Challenge, vol. 5, p. 12 (2015)

[24] Wasserthal, J., Meyer, M., Breit, H., Cyriac, J., Yang, S., Segeroth, M.: Totalsegmentator: Robust segmentation of 104 anatomical structures in ct images 2022. arXiv (2022)

[25] 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)

[26] He, Y., Yang, G., Yang, J., Ge, R., Kong, Y., Zhu, X., Zhang, S., Shao, P., Shu, H., Dillenseger, J.-L., *et al.*: Meta grayscale adaptive network for 3d integrated renal structures segmentation. Medical image analysis **71**, 102055 (2021)

[27] He, Y., Yang, G., Yang, J., Chen, Y., Kong, Y., Wu, J., Tang, L., Zhu, X., Dillenseger, J.-L., Shao, P., *et al.*: Dense biased networks with deep priori anatomy and hard region adaptation: Semi-supervised learning for fine renal artery segmentation. Medical image analysis **63**, 101722 (2020)

[28] Shao, P., Qin, C., Yin, C., Meng, X., Ju, X., Li, J., Lv, Q., Zhang, W., Xu, Z.: Laparoscopic partial nephrectomy with segmental renal artery clamping: technique and clinical outcomes. European urology **59**(5), 849–855 (2011)

[29] Shao, P., Tang, L., Li, P., Xu, Y., Qin, C., Cao, Q., Ju, X., Meng, X., Lv, Q., Li, J., *et al.*: Precise segmental renal artery clamping under the guidance of dual-source computed tomography angiography during laparoscopic partial nephrectomy. European urology **62**(6), 1001–1008 (2012)

[30] Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International journal of computer vision **59**, 167–181 (2004)

[31] brgfx: Image by brgfx on Freepik. https://www.freepik.com/free-vector/anatomical-structure-human-body 27539420.htm

[32] 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)

[33] Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment Anything in Medical Images (2023)

[34] Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sam-med2d. arXiv preprint arXiv:2308.16184 (2023)

[35] Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., Fu, B., Zhang, S., He, J., Qiao, Y.: SAM-Med3D (2023)

[36] Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. arXiv preprint arXiv:2206.08023 (2022)

[37] Grauw, M.: Universal Lesion Segmentation Challenge 23. https://uls23.grand-challenge.org/

[38] Kavur, A.E., Gezer, N.S., Bar, M., Aslan, S., Conze, P.-H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., zkan, S., Baydar, B., Lachinov, D., Han, S., Pauli, J., Isensee, F., Perkonigg, M., Sathish, R., Rajan, R., Sheet, D., Dovletov, G., Speck, O., Nrnberger, A., Maier-Hein, K.H., Bozda Akar, G., nal, G., Dicle, O., Selver, M.A.: Chaos challenge - combined (ct-mr) healthy abdominal organ segmentation. Medical Image Analysis **69**, 101950 (2021) https://doi.org/10.1016/j.media.2020.101950

[39] Kavur, A.E., Selver, M.A., Dicle, O., Bar, M., Gezer, N.S.: Chaos - combined (ct-mr) healthy abdominal organ segmentation challenge data (2019) https://doi.org/10.5281/zenodo.3362844

[40] Kavur, A.E., Gezer, N.S., Bar, M., ahin, Y., zkan, S., Baydar, B., Yksel, U., Klker, ., Olut, ., Bozda Akar, G., nal, G., Dicle, O., Selver, M.A.: Comparison of semi-automatic and deep learning based automatic methods for liver segmentation in living liver transplant donors. Diagnostic and Interventional Radiology **26**, 11–21 (2020) https://doi.org/10.5152/dir.2019.19

[41] 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) https://doi.org/10.1109/TPAMI.2021.3100536

[42] 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)

[43] 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, 101821 (2020)

[44] Roth, H.R., Farag, A., Turkbey, E., Lu, L., Liu, J., Summers, R.M.: Data from pancreas-ct. the cancer imaging archive. IEEE Transactions on Image Processing (2016)

[45] Roth, H.R., Lu, L., Farag, A., Shin, H.-C., Liu, J., Turkbey, E.B., Summers, R.M.: Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany,

October 5-9, 2015, Proceedings, Part I 18, pp. 556–564 (2015). Springer

[46] Soler, L., Hostettler, A., Agnus, V., Charnoz, A., Fasquel, J.-B., Moreau, J., Osswald, A.-B., Bouhadjar, M., Marescaux, J.: 3d image reconstruction for comparison of algorithm database. URL: https://www. ircad. fr/research/data-sets/liver-segmentation-3d-ircadb-01 (2010)

[47] Heimann, T., Van Ginneken, B., Styner, M.A., Arzhaeva, Y., Aurich, V., Bauer, C., Beck, A., Becker, C., Beichel, R., Bekes, G., *et al.*: Comparison and evaluation of methods for liver segmentation from ct datasets. IEEE transactions on medical imaging **28**(8), 1251–1265 (2009)

[48] Quantification of Uncertainties in Biomedical Image Quantification Challenge 2021. https://qubiq21.grand-challenge.org/. Accessed: 18 Aug 2023

[49] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (2021)

[50] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[51] He, K., Chen, X., Xie, S., Li, Y., Dollr, P., Girshick, R.: Masked Autoencoders Are Scalable Vision Learners (2021)

[52] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision (2021)

[53] Liu, J., Zhang, Y., Chen, J.-N., Xiao, J., Lu, Y., Landman, B.A., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection (2023)

[54] Tancik, M., Srinivasan, P.P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J.T., Ng, R.: Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains (2020)

[55] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need (2023)