Journal of Imaging Informatics in Medicine (2024) 37:1529–1547
https://doi.org/10.1007/s10278-024-00981-7

From CNN to Transformer: A Review of Medical Image Segmentation Models

Wenjian Yao[1] · Jiajun Bai[1] · Wei Liao[2] · Yuheng Chen[1] · Mengjuan Liu[1] · Yao Xie[3,4]

Received: 10 July 2023 / Revised: 13 November 2023 / Accepted: 14 November 2023 / Published online: 4 March 2024

Abstract

Medical image segmentation is an important step in medical image analysis, especially as a crucial prerequisite for efficient disease diagnosis and treatment. The use of deep learning for image segmentation has become a prevalent trend. The widely adopted approach currently is U-Net and its variants. Moreover, with the remarkable success of pre-trained models in natural language processing tasks, transformer-based models like TransUNet have achieved desirable performance on multiple medical image segmentation datasets. Recently, the Segment Anything Model (SAM) and its variants have also been attempted for medical image segmentation. In this paper, we conduct a survey of the most representative seven medical image segmentation models in recent years. We theoretically analyze the characteristics of these models and quantitatively evaluate their performance on Tuberculosis Chest X-rays, Ovarian Tumors, and Liver Segmentation datasets. Finally, we discuss the main challenges and future trends in medical image segmentation. Our work can assist researchers in the related field to quickly establish medical segmentation models tailored to specific regions.

Keywords Deep learning · Medical image segmentation · CNN · U-Net · Transformer

Introduction

With the continuous development of medical imaging technology, medical images have become essential for disease diagnosis and treatment planning [1]. Medical image segmentation plays a vital role among the foundational and critical techniques in medical image analysis. Medical image segmentation refers to the identification of organ or lesion pixels from medical images such as CT or MRI [2, 3]. We believe that it is an important step in medical image analysis, aiming to convey and extract crucial information about the shape and volume of these organs or tissues. Traditional methods for medical image segmentation primarily rely on manual feature extraction by physicians or handcrafted designs based on image processing techniques and mathematical models, such as thresholding [4], edge detection [5], and morphological operations. These methods offer a certain level of interpretability and controllability. However, due to the complexity and diversity of medical images, as well as the specificity of medical image segmentation tasks, traditional segmentation methods have certain limitations. Handcrafted algorithms fail to meet the requirements of efficiency and accuracy when dealing with a large number of medical images for segmentation tasks. Moreover, manual

✉ Mengjuan Liu
mjliu@uestc.edu.cn

✉ Yao Xie
xieyao@med.uestc.edu.cn

Wenjian Yao
1304319616@qq.com

Jiajun Bai
haku_bill@outlook.com

Wei Liao
770147659@qq.com

Yuheng Chen
823368720@qq.com

[1] Network and Data Security Key Laboratory of Sichuan Province, University of Electronic Science and Technology of China, 610054 Chengdu, China

[2] Department of Obstetrics and Gynaecology, Deyang People's Hospital, 618000 Deyang, China

[3] Department of Obstetrics and Gynaecology, Sichuan Provincial People's Hospital, School of Medicine, University of Electronic Science and Technology of China, Chengdu, China

[4] Chinese Academy of Sciences Sichuan Translational Medicine Research Hospital, 610072 Chengdu, China

feature extraction from medical images requires physicians with rich expertise and experience, making them susceptible to subjective factors.

Deep learning techniques have been widely applied in medical image segmentation in recent years to address the issues above. Through deep feature learning, models can extract semantic information from images, thereby improving segmentation accuracy and flexibly adapting to different medical image datasets and tasks. Segmentation models based on fully convolutional networks (FCNs) and convolutional neural networks (CNNs) have achieved remarkable results. For example, the U-Net model won first place in the ISBI 2015 Cell Segmentation Challenge [6], and the SegNet model demonstrated good performance in semantic segmentation tasks on the CamVid dataset [7], among others. After this, Google proposed DeepLab [8], a series of semantic segmentation models based on dilated convolutions. However, convolutional neural networks have limited modeling capabilities for long-range dependencies, making it challenging to exploit the semantic information within images fully.

Recently, some new segmentation models have been proposed, including TransUNet [9] and Swin-Unet [10]. TransUNet is a segmentation model that introduces Transformer modules [11] to improve the model's ability to model long-range dependencies. The Transformer module adopts a self-attention mechanism, which calculates the similarity between each position and other positions in the input sequence, resulting in a weight vector. This weight vector is used to compute weighted representations for each position, facilitating the interaction and integration of global information. In other words, the Transformer model can effectively capture the correlations between different positions in the input sequence through the self-attention mechanism, thereby better understanding and processing sequential data. In TransUnet, the Transformer module is embedded within a U-shaped architecture to extract global information from the image, enhancing the model's semantic representation capability and making it more suitable for handling large-sized, high-resolution medical images.

Swin-Unet, on the other hand, is another novel segmentation model that introduces the Swin Transformer module [12] to improve computational efficiency. The Swin Transformer is a hierarchical self-attention mechanism that decomposes the input feature map into multiple small patches, with each patch independently computing attention weights, thus reducing computational complexity. The Swin Transformer module in Swin-Unet is combined with a U-shaped architecture, allowing for the extraction of global information from the image while reducing computational complexity and memory consumption. This makes it more suitable for medical image segmentation tasks.

Image segmentation based on machine learning has developed rapidly. Meta [13] proposed the Segment Anything Model (SAM), revolutionizing image segmentation. It is the first time to introduce the concept of foundation models in image segmentation with zero-shot migration. Unlike previous image segmentation models that can only handle a particular class of images, SAM can handle all images and achieve accurate image segmentation by the prompt.

Despite the vibrant development of medical image segmentation techniques in recent years, there is still a lack of comprehensive review papers on the application of deep learning models in medical image segmentation, particularly the introduction of the latest segmentation models and quantitative performance comparisons among these models. The literature [14, 15] covers only traditional CNN-based segmentation models, while literature [16] focuses solely on the model structure without quantitative evaluation. This paper conducts a survey on the seven most representative medical image segmentation models in recent years: FCN, U-Net, DeepLab, UNet++ [17], TransUNet, Swin-Unet, and SAM. The characteristics of these models are analyzed theoretically, and their performance is quantitatively evaluated on three benchmark datasets. Finally, we discuss the main challenges and future development trends in medical image segmentation. Furthermore, we have shared all experimental source code and detailed model configuration parameters on GitHub to assist related researchers in quickly understanding these models and modeling new segmentation tasks.

The rest of this paper is organized as follows. The "Typical Medical Image Segmentation Models" section describes representative medical image segmentation methods. The "Experimental Setup" section introduces the datasets and provides relevant experimental details. The "Experimental Results" section presents the evaluation results of each model conducted in our study. Finally, the paper concludes with an outlook on the challenges and future developments in medical image segmentation.

Typical Medical Image Segmentation Models

Deep learning has significantly advanced medical image segmentation in recent years [18]. Convolutional neural networks (CNNs) [19], especially fully convolutional networks (FCNs) [20], dominate medical image segmentation. With the evolution of medical image segmentation, among different model variants, U-Net has become the de facto choice, consisting of a symmetric encoder-decoder network with skip connections for improved detail retention. Image features can be automatically extracted and used in segmentation tasks by neural networks. This section presents seven representative models

that have performed particularly well in recent years, including FCN, U-Net, DeepLab, UNet++, TransUNet, Swin-Unet, and SAM.

FCN

Fully convolution network (FCN) was first proposed in 2014 [20]. In CNN, the fully connected layer usually follows the final convolution layer to map the feature map produced by the convolution layer into a fixed-length vector. Thus, CNN is suitable for image-level classification and regression tasks since both expect the probability of classification of the input image at the end. FCN is a fully convolutional network without fully connected layers, which classifies the image at the pixel level and solves the semantic segmentation problem of the image. FCN can accept input images of any size and upsample the feature map produced by the last convolutional layer to return it to the size of the input image. In this way, FCN is able to perform more accurate pixel-by-pixel classification.

Depending on the different strides of the deconvolutions, there are different versions, FCN-32 s, FCN-16 s, and FCN-8 s. FCN-32 s directly samples the input image size by one deconvolution after the fifth convolution, FCN-16 s conducts two deconvolutions, and fuses the result of the fourth convolution with one deconvolution, and the stride of the second deconvolution is 16. FCN-8 s dose three deconvolutions. Similarly, it further fuses the result of the third convolution and the stride of the third deconvolution is 8. The experimental results of the original paper show that FCN-8 s has the best segmentation results, which indicates that the prediction results of the shallower layers contain more detailed features, which helps in a more accurate segmentation, so FCN-8 s is used for the experiments in this paper. The model structure is depicted in Fig. 1.

U-Net

U-Net is one of the most well-known network architectures in the medical image segmentation models. It was proposed by Ronneberger et al. [6] for the ISBI Challenge in 2015. The U-Net model is considered a classic model in medical image segmentation and has been widely applied to various tasks, including CT, MRI, and X-ray segmentation. The model structure is depicted in Fig. 2. Its success lies in combining the deep feature extraction capability of convolutional neural networks (CNNs) with the pixel-level segmentation ability of fully convolutional networks (FCNs). It also incorporates techniques such as skip connections to leverage both low-level and high-level feature information, thereby improving segmentation accuracy and robustness.

The U-Net model consists of a contracting path and an expanding path. The contracting path follows a typical architecture of convolutional networks. In each downsampling step, the number of feature channels is doubled. Each step in the expanding path includes upsampling feature maps, halving the number of feature channels, and concatenating them with the correspondingly cropped feature maps from the contracting path. In the final layer, a 1x1 convolution is applied to map each 64-component feature vector to the desired number of classes. The network consists of a total of 23 convolutional layers.

Since the introduction of the U-Net model, several improved versions based on U-Net have emerged, including UNet++, Attention-UNet, TransUNet, and Swin-Unet, among others. These models build upon the advantages of the original U-Net model and further enhance segmentation performance by introducing attention mechanisms,

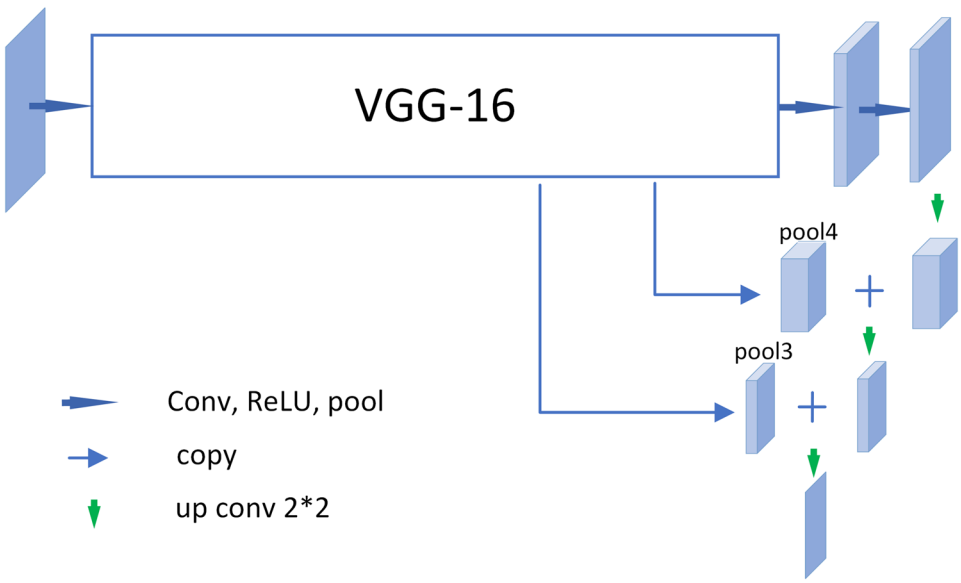

Fig. 1 Structure of FCN-8 s [20]

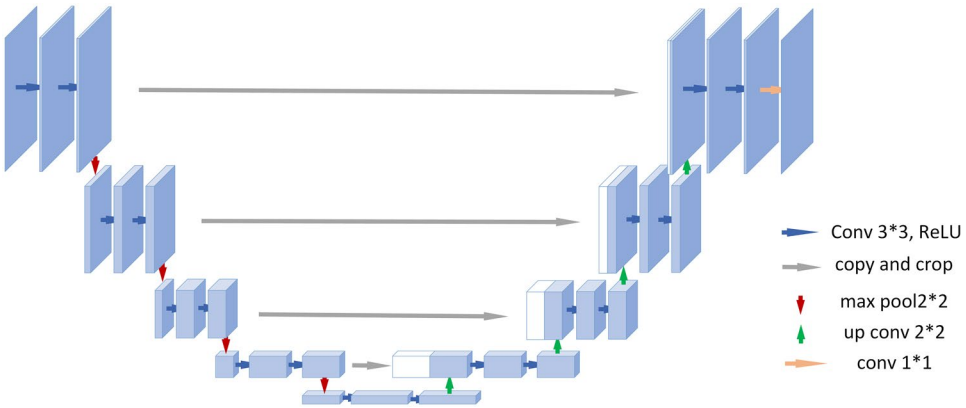

Fig. 2 Structure of U-Net [6]

transformational network structures, and other techniques. As a result, the U-Net model holds an important position and influence in medical image segmentation.

DeepLab

DeepLab has evolved between 2015 and 2018, with a total of four releases called V1, V2 [21], V3 [22], and V3+ [23]. Same as FCN, DeepLab V1 uses VGG-16 as a backbone network to extract features and classify images at the pixel level. Based on this, DeepLab V1 introduces atrous convolution. Atrous convolution can increase the sensory field of view (FOV) without changing the size of the output feature map of the image, and at the same time, the convolution kernel expansion of atrous convolution does not increase the computational effort of the model. This makes it feasible to obtain multi-scale contextual information by cheaply stacking multiple void convolution layers with different expansion rates [24].

In DeepLab V2, Backbone is switched from VGG-16 to ResNet-101. DeepLab V2 proposes the Atrous Spatial Pyramid Pooling (ASPP) module, which improves the SPP module [25] in an atrous way. Meanwhile, DeepLab V2 implements multi-scale feature fusion by running the original image in parallel three times on reduced images with scales of 1.0, 0.75, and 0.5 and taking the maximum value of the result. DeepLab V3 further improves the ASPP module by making the module capable of extracting the global features of a single image as well as the tiny features of local regions. A batch normalization (BN) layer is added between the final convolution and ReLU of the ASPP module, which improves the training. DeepLab V3+ finally adopts the encoder-decoder structure to further optimize the model performance based on DeepLab V3, while Xecption is adopted as the backbone. In the subsequent experiments, we use DeepLab V3+ (Resnet-101) and DeepLab V3+ (Xecption) as benchmark models.

UNet++

The UNet++ network architecture was proposed by Zhou et al. in 2018 [17], introducing the concept of dense connections to the U-Net network. The model structure is depicted in Fig. 3. UNet++ builds upon the U-Net model while retaining the long skip connections and adds more short-skip connection paths and upsampling convolution blocks to form new levels of encoders. The U-shaped connectivity structure in UNet++ is achieved by fusing each encoder in the decoder with other encoders at the same level. Specifically, each encoder receives feature maps of the same scale from other encoders and concatenates them to obtain more discriminative feature representations. Furthermore, the later proposed attention-UNet++ [26] improves the feature map concatenation by adding attention mechanisms to enhance the focus and extraction of important features during encoder fusion.

UNet++ captures features from different levels by introducing dense connections, enabling the extraction of feature information from different layers and scales. These features are integrated into the final prediction to improve segmentation accuracy. The idea of dense connections originated from DenseNet [27]. Prior to DenseNet, the evolution of convolutional neural networks typically involved increasing the depth or width of the networks. DenseNet introduced a new structure by reusing features, which not only alleviated the problem of gradient vanishing but also reduced the number of model parameters. In the original U-Net network architecture, the use of depth supervision in the intermediate hidden layers addresses the issue of gradient vanishing during UNet++ training. It also allows for network pruning during the testing phase, reducing the inference time of the model.

TransUNet

TransUNet network architecture was proposed by Chen et al. in 2021 as a Transformer-based segmentation network

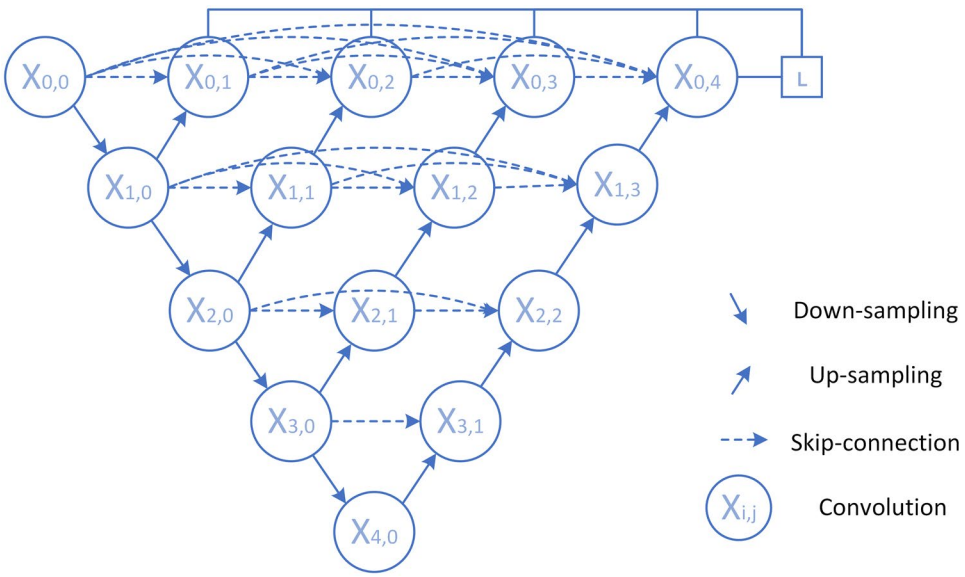

Fig. 3 Structure of UNet++ [17]

[9]. The model structure is depicted in Fig. 4. TransUNet builds upon the U-Net model by introducing a hybrid encoder that combines CNN and Transformer to address the limitations of traditional convolutional neural networks in modeling long-range dependencies and handling large-sized images. The core of TransUNet is the Transformer module [11], which consists of multi-head self-attention mechanisms and feed-forward neural networks. The multi-head self-attention mechanism captures dependencies between different positions in the image, establishing global contextual information in the feature representation. This enables TransUNet to handle long-range dependencies better, capture semantic information in the image, and improve the model's representation capacity and generalization performance.

Specifically, TransUNet first uses CNN to extract features and generate feature maps of the input image. These feature maps are then divided into patches of size 1x1 and fed into an additional stack of 12 Transformer modules. This hybrid structure combines convolutional neural networks' feature extraction capability with effective global information modeling using Transformer modules, yielding better performance than using pure Transformers as encoders. The decoder in TransUNet performs upsampling on the encoded features and combines them with high-resolution CNN feature maps to enrich the semantic information, achieving more precise localization. The final step involves restoring the feature maps to the original image size and generating pixel-level segmentation results. Compared to the traditional U-shaped models that use convolutional neural networks, TransUNet introduces a stack of 12 Transformer modules, significantly increasing the number of parameters. It increases the difficulty of model training. In this study, a suboptimal approach of reducing the batch size was employed to meet the training requirements of TransUNet on GPUs.

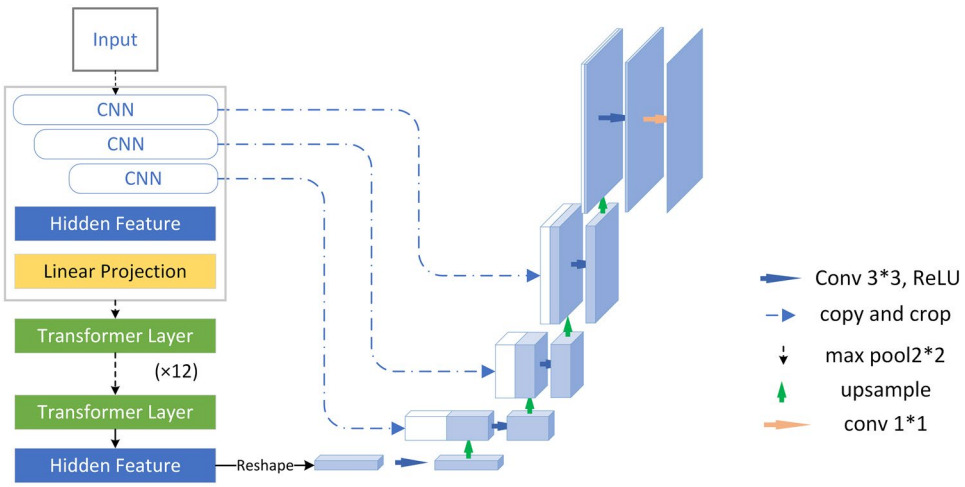

Fig. 4 Structure of TransUNet [9]

Swin-Unet

Swin-Unet network architecture was proposed by Cao et al. in 2023 [10]. The model structure is depicted in Fig. 5. Unlike TransUnet, which replaces the convolutional blocks in the U-Net encoder with Transformer blocks, Swin-Unet utilizes Swin Transformer blocks [12] to extract hierarchical features from the input image. Swin-Unet is the first purely Transformer-based U-shaped architecture. Swin Transformer extends the traditional Transformer's one-dimensional sequence to two-dimensional image blocks and adopts a hierarchical attention mechanism to capture features over a larger receptive field. This structure is similar to the hierarchical structure in convolutional neural networks and serves feature extraction. Additionally, the Swin Transformer introduces the mechanism of shifting windows on top of the self-attention mechanism. By limiting the attention calculation to windows in the vicinity of the current region, Swin-Unet better preserves positional information and further improves the model's performance.

In Swin-Unet, Swin Transformer is applied in the encoding, bottleneck, and decoding modules. Importantly, the compression of each layer's features in Swin-Unet is smaller than in TransUNet. Instead of adding additional Transformer modules, Swin-Unet replaces the convolutional modules with Transformer modules, effectively reducing the number of model parameters. Overall, Swin-Unet leverages the advantages of Swin Transformer and U-Net to provide a promising approach for medical image segmentation. It has demonstrated competitive performance in various segmentation challenges and benchmarks.

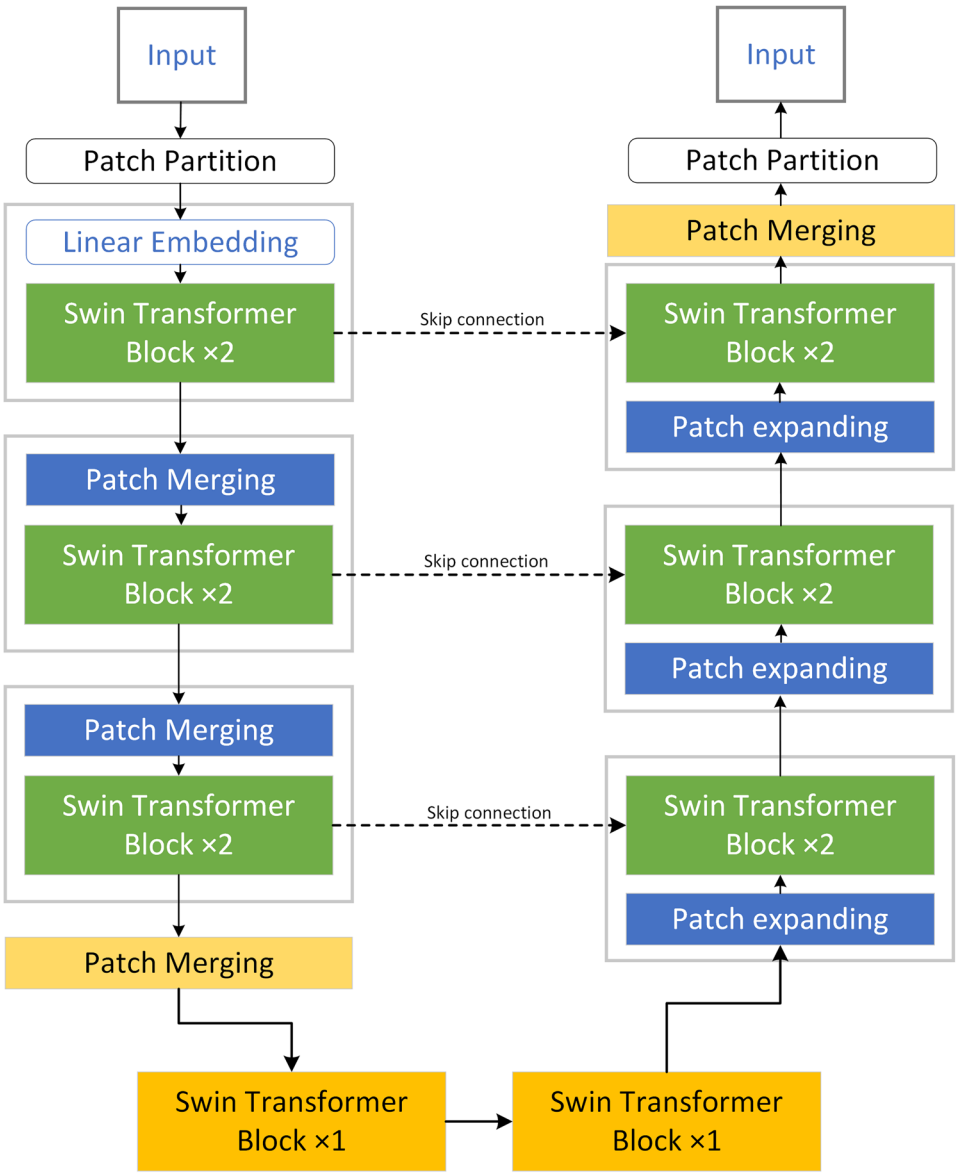

Fig. 5 Structure of Swin-Unet [10]

Segment Anything Model

Meta [13] proposed SAM in 2023, which uses prompt techniques to perform zero-shot learning and few-shot learning on new datasets and tasks. Meta has collected the largest segmentation dataset ever, the Segment Anything 1-Billion mask dataset (SA-1B), which contains 11 million images and over 1 billion mask maps in total. The model is designed to be interactive and promptable during training so that it can be transferred to new image distributions and tasks through zero-shot or few-shot learning.

SAM consists of three main parts: image encoder, prompt encoder, and mask decoder. The image encoder is a pre-trained Vision Transformer (ViT), where the masked autoencoder (MAE) is used to extract the image embedding of high-resolution images [28]. Masked prompts inputted to the model are classified into two types: sparse prompts (point, box, and text) and dense prompts (mask). The point prompt is the position encoding as well as the foreground/background identifiers, the box prompt is the position encoding representing the upper left and lower right position of the box, and CLIP [29] is used to generate text encoding for text prompts. The sparse prompts produced by the prompt encoder are directly involved in the computation of the attention layer as prompt tokens. Unlike sparse prompts, the mask prompt is first convolved and then summed element-wise with image embedding, by which the mask features are embedded in the image. The role of the mask decoder is to map the image embedding and one of the above classes of prompt embedding to an output mask. The structure of the mask decoder is modified from a standard Transformer decoder [30]. A set of vectors is obtained after passing the output token of the mask decoder to a three-layer MLP. The output image embedding is upsampled and then combined with vectors to get a set of prediction masks for the image. At the same time, the model outputs a set of IoU scores corresponding to the degree of matching of the above set of masks to meet the different needs of downstream tasks.

Experimental Setup

Datasets

This subsection presents the datasets used for evaluating the seven typical segmentation models.

- **Tuberculosis Chest X-rays dataset** [31]. The dataset was acquired from the Department of Health and Human Services, Montgomery County, MD, USA, and Shenzhen No. 3 People's Hospital in China. The chest X-rays were from outpatient clinics and were captured as part of the daily hospital routine within one month, mostly in September 2012, using a Philips DR Digital Diagnose system. The dataset contains normal cases and cases with TB manifestations. In the experiments of this paper, we do not need to consider whether the samples have TB manifestation but focus on the segmentation effect of the model on the lung region. We select 566 clear chest x-ray images to form the experimental dataset and randomly divide the dataset into 452 training samples and 114 test samples in a ratio of 4:1. Some of the images and labels are shown in Fig. 6.

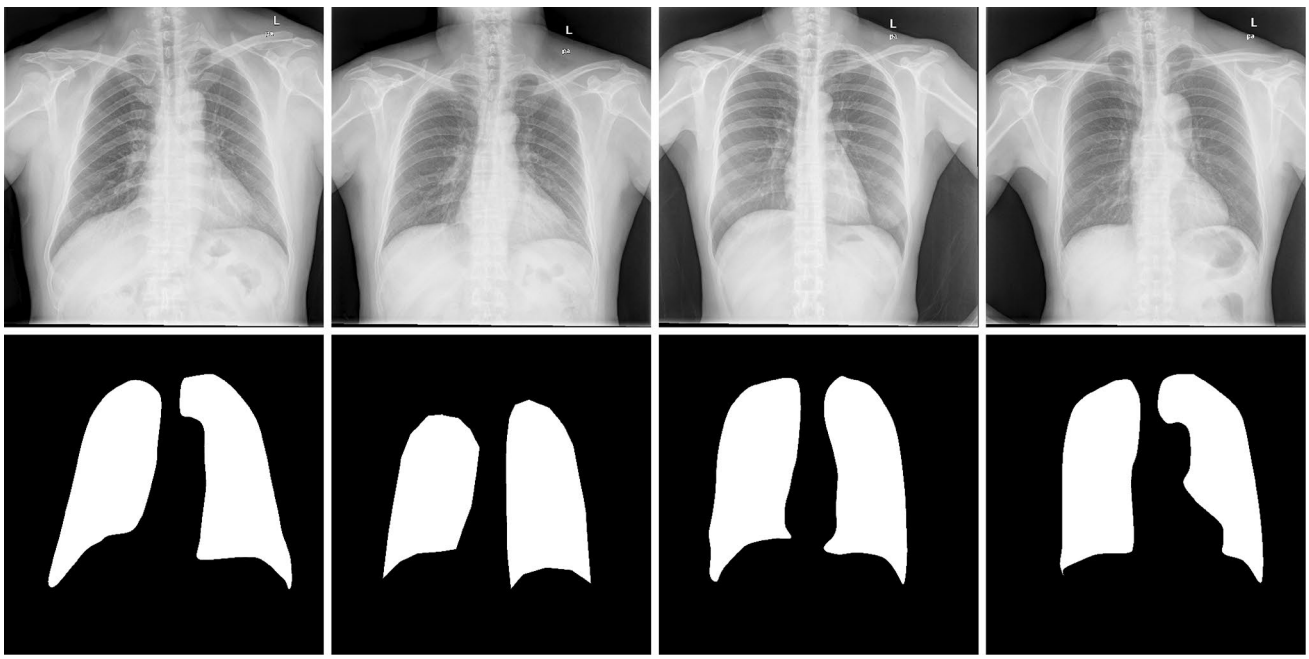

Fig. 6 Examples of Tuberculosis Chest X-rays (Shenzhen)dataset

- **Clinical Liver CT dataset**. The dataset is from Segmentation of the Liver Competition 2007 (SLIVER07) [32]. The segmentation goal of this dataset is to segment the liver region from clinical 3D Computed Tomography scans. We perform a certain degree of preprocessing on the dataset, including adjusting the window width to 40 and the window position to 400, slicing the 3D structures, and removing slices that do not contain the target area and those that are too small. Finally, 2421 different images are obtained as the experimental dataset. Then, we divide the dataset into 1937 training samples and 484 test samples in a ratio of 4:1. Examples of images and labels are shown in Fig. 7.
- **Ovarian Tumors dataset**. We collected a 2D CT segmentation dataset of abdominopelvic ovarian masses, which contains CT images of the patients as well as segmentation labels. The entire dataset was selected and annotated by physicians specializing in obstetrics and gynecology at the People's Hospital of Sichuan Province, China. The pelvic CT images of 123 patients were collected using Computed Tomography, including 35 patients with benign tumors, 83 with malignant tumors, and 5 with unknown tumor benignity or malignancy. Since the CT images are three-dimensional, we need to slice the CT images of each patient horizontally and select several two-dimensional images containing the masses to construct the sample set. Based on this method, we constructed a CT image dataset containing 4050 abdominopelvic CT images of 123 patients. Next, we randomly selected 3092 CT images from 98 patients as the training set and 958 images from 25 patients as the test set. Such a construction method ensures that different images of the same patient will not appear in the training and test sets at the same time so that the segmentation effect of the model will not be affected by the image pattern of the same patient. Some images and labels are shown in Fig. 8.

Implementation Details

This subsection introduces the loss function and training parameters used in our experiments.

- **Loss function**. The loss functions used in our work are binary cross-entropy and Dice coefficient. Binary cross-entropy is utilized to evaluate the performance of the binary classification model, where we consider the regions of interest and non-interest in the images as 1 and 0, respectively. The formula for binary cross-entropy loss is described as Formula (1), where y represents the actual labels of each pixel in the image and $\hat{y}$ is the predicted value of each pixel determined by the model. On the other hand, the Dice coefficient is employed to measure the overlap between the predicted image and the ground truth image. The formula of Dice loss is described as Formula (2), in which TP, FP, TN, and FN respectively represent the predicted true positive, false positive, true negative, and false negative in the confusion matrix, as shown in Table 1 in image segmentation.

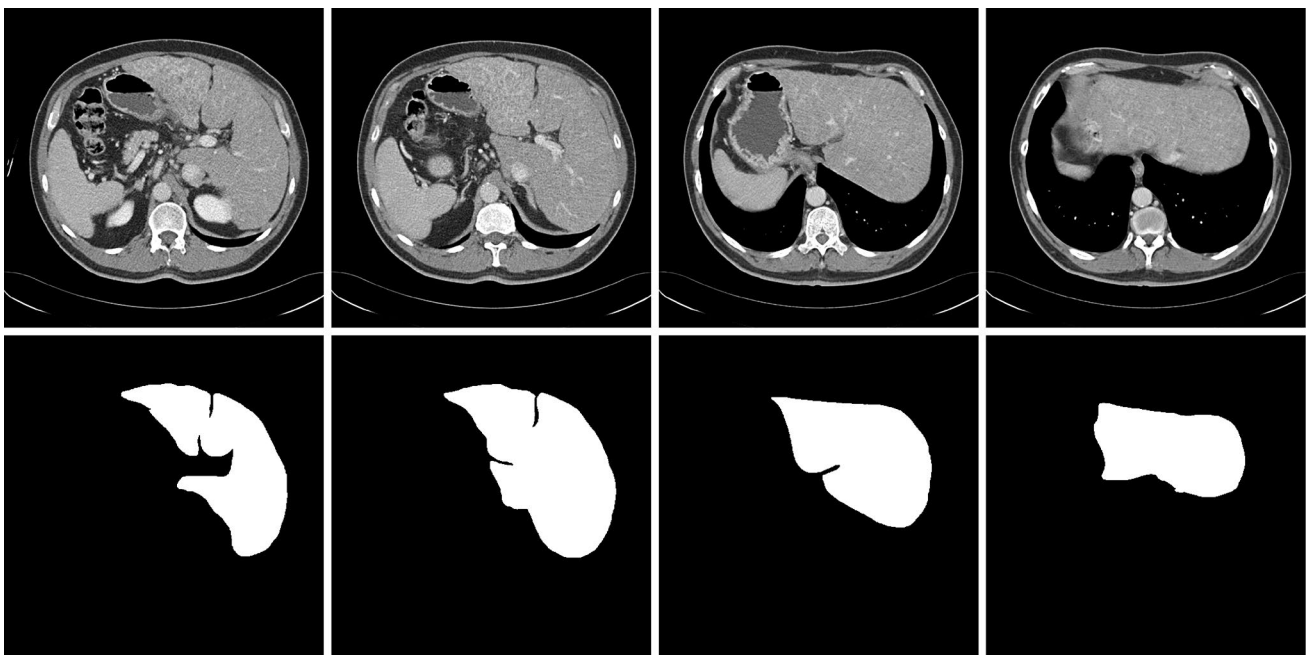

Fig. 7 Examples of Clinical Liver CT dataset

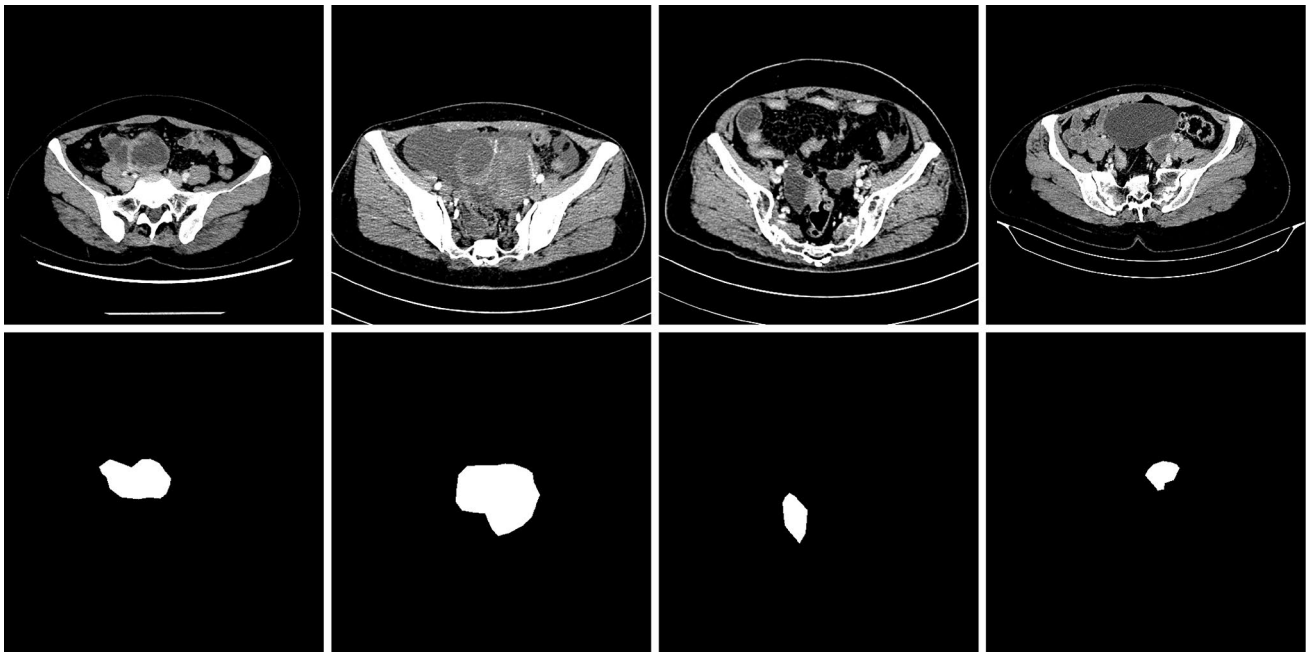

Fig. 8 Examples of Ovarian Tumors dataset

$$L_{BCE} = -[(1-y)log(1-\hat{y}) + ylog\hat{y}] \quad (1)$$

$$L_{DSC} = 1 - \frac{2TP}{2TP + FP + FN} \quad (2)$$

- **Training parameters**. The experiments are trained and tested based on Python and Pytorch implementations. For the FCN-8 s model, the input size is set to 512 × 512, SGD optimizer is used with a learning rate of 1e-3, a momentum of 0.9, and a weight decay of 5e-4 for backpropagation. For U-Net and UNet++, the input size is set to 512 × 512. The models are trained using the Adam optimizer with a learning rate of 1e-3 and weight decay of 1e-8 for backpropagation. For DeepLab models, the input size is set to 512 × 512. The models are trained using the SGD optimizer with a learning rate of 1e-3 a momentum of 0.9 and weight decay of 5e-4 for backpropagation. For the TransUNet model, the input size and patch size are set to 512 × 512 and 16, respectively. The Transformer backbone and ResNet-50 utilize pre-trained weights from ImageNet. The model is optimized using the SGD optimizer with a momentum of 0.9, and weight decay of 1e-4 for backpropagation. For the Swin-Unet model, the input size and patch size are set to 224 × 224 and 4, respectively. Pre-trained weights from ImageNet, provided by the original authors, are used for initializing the model parameters. The model is optimized using the SGD optimizer with a momentum of 0.9 and weight decay of 1e-4 for backpropagation. Table 2 shows the parameter settings for each model. Additionally, early stopping is introduced in the training of all models to prevent overfitting.

In addition, for SAM, we validate the segmentation performance directly on the test set. Prompts are generated with the following strategy: for point prompts, we randomly generate the coordinates of points in the

Table 1 Confusion matrix in image segmentation

		Predicted sample	
		Positive	Negative
Actual Sample	Positive	TP (Pixels belong to real regions of interest and are predicted as regions of interest)	FP (Pixels belong to real regions of interest but are predicted as regions of no interest)
	Negative	FN (Pixels belong to real regions of no interest and are predicted as regions of interest)	TN (Pixels belong to real regions of no interest and are predicted as regions of no interest)

Table 2 parameter settings for each model

	Input_Size	Patch_Size	Learning_Rate	Optimizer
FCN	512*512	1	1e-3	SGD
U-Net	512*512	1	1e-3	Adam
DeepLabs	512*512	1	1e-3	SGD
UNet++	512*512	1	1e-3	Adam
TransUNet	512*512	16	1e-2	SGD
Swin-UNet	224*224	4	1e-2	SGD

foreground of ROIs in the test image to simulate human mouse clicks; for box prompts, we generate the smallest rectangle that can encompass all ROIs in the image to simulate human box selection.

Experimental Results

Evaluation Metrics

We employ the following medical image segmentation evaluation metrics to assess the performance of the model. Additionally, we observe the actual segmentation results to comprehensively evaluate the segmentation capability of the model.

- **Dice**: The segmentation capability of a segmentation model is commonly measured using the Dice coefficient, which represents the similarity between two samples. It has a value range of [0, 1], where a higher value indicates better model performance. The Dice loss is described by Formula (2).
- **HD95**: HD95 refers to the 95% Hausdorff distance, which quantifies the maximum distance between two sets at the 95th percentile. A smaller value indicates a higher similarity between the two sets. The HD95 is described by Formula (3), where h(A, B) denotes the maximum value in the shortest distance to each pixel in B calculated for each pixel in A.

$$\begin{aligned} HD &= max_{k95\%}(h(A,B),h(B,A)) \\ h(A,B) &= max(a \in A)min(b \in B)||a-b|| \\ h(B,A) &= max(b \in B)min(a \in A)||a-b|| \end{aligned} \tag{3}$$

- **IoU**: The IoU (Intersection over Union) score is a standard performance measure for image segmentation problems. It also measures the similarity between two samples. The IoU score is described by Formula (4).

$$IoU = \frac{TP}{TP+FP+FN} \tag{4}$$

- **Accuracy**: Accuracy is one of the commonly used metrics to measure the segmentation capability of a model. It specifically calculates the ratio of correctly classified pixels to the total number of pixels in the dataset. The accuracy is described by Formula (5).

$$Acc = \frac{TP+TN}{TP+FN+FP+TN} \tag{5}$$

- **Precision**: Precision, also known as the positive predictive value, represents the ratio of true positive samples to the total predicted positive samples by the model. A higher precision value, closer to 1, is desired. Precision is described by Formula (6).

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

- **Recall**: Recall, also known as sensitivity or true positive rate, represents the ratio of true positive samples to the total actual positive samples. A higher recall value, closer to 1, is desirable. Recall is described by Formula (7).

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

Results Analysis

We first evaluated the performance of each model on a publicly available dataset, the Tuberculosis Chest X-rays dataset. Table 3 lists the experimental results for each

Table 3 Tuberculosis Chest X-rays (Shenzhen) dataset test index

	DSC↑	HD95↓	IoU↑	Acc↑	Precision↑	Recall↑
FCN	92.89%	19.60	87.02%	96.47%	93.56%	0.9269%
U-Net	95.32%	14.23	91.24%	97.69%	96.13%	94.78%
DeepLab V3+(Resnet101)	95.17%	12.63	90.87%	97.54%	94.23%	**96.31%**
DeepLab V3+(Xception)	90.37%	48.19	82.85%	94.94%	86.19%	95.91%
UNet++	95.83%	11.75	92.15%	97.95%	97.31%	94.62%
TransUNet	**96.45%**	**10.75**	**93.25%**	**98.16%**	**97.36%**	95.72%
Swin-Unet	95.71%	12.10	91.88%	97.80%	96.81%	94.79%
SAM-1point	41.79%	180.61	30.01%	65.82%	50.15%	49.17%
SAM-box	90.14%	28.56	82.40%	94.57%	90.81%	90.43%

Bolded datas represent the optimal value of each model for this indicator

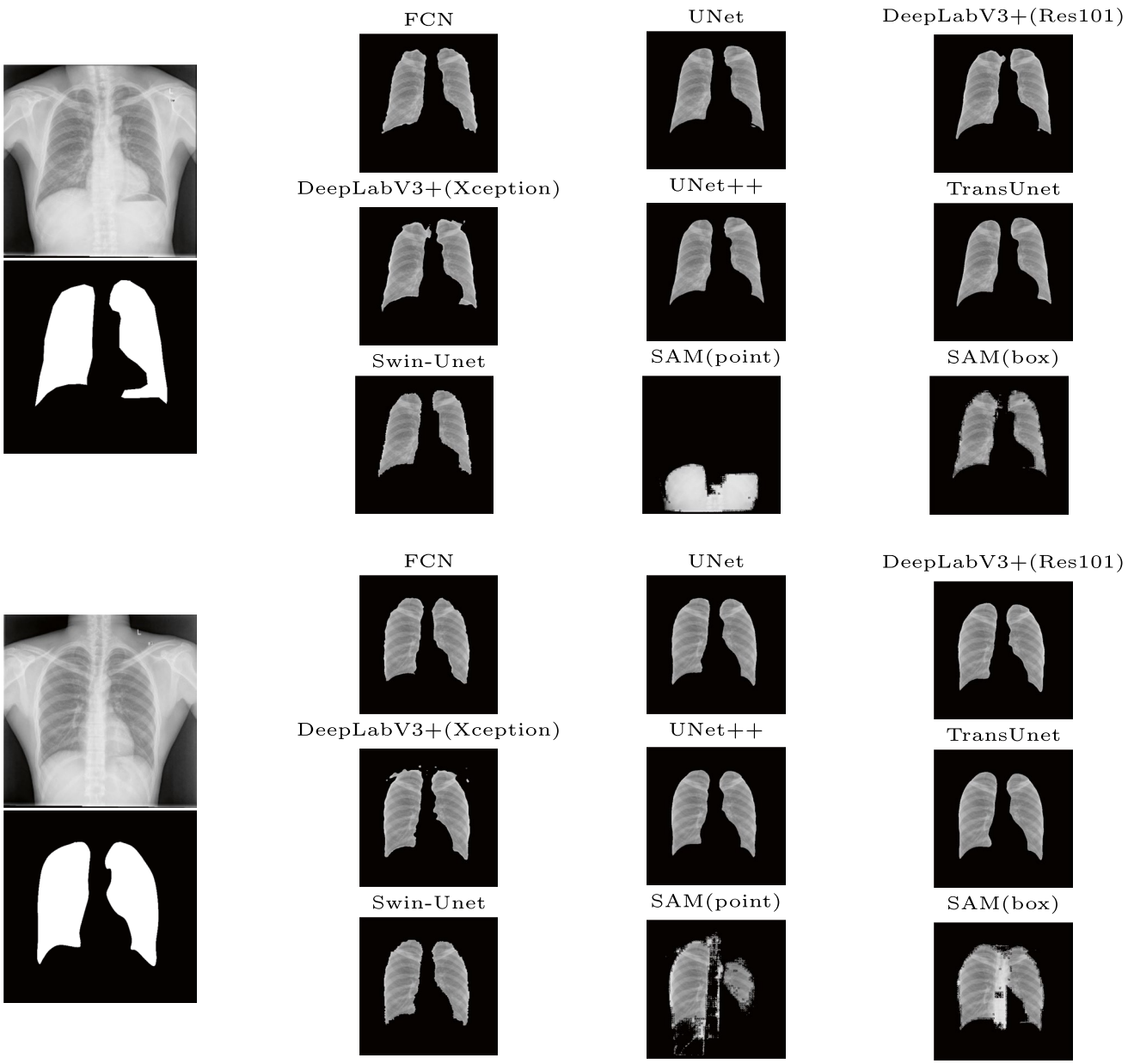

Fig. 9 Tuberculosis Chest X-rays (Shenzhen)dataset segmentation results

model, while Fig. 9 visualizes the model performance. Figure 10 visualizes the coloring result where the green area represents tp, red area represents fp and yellow area represents fn. The results show that the TransU-Net model achieves the best performance in five of the metrics, 96.45% (DSC↑), 10.75 (HD↓), 93.25% (IoU↑), 98.16% (Acc↑), and 97.36% (Precision↑), while DeepLab V3+ (Resnet101) achieves the best performance in the remaining the best recall metrics (96.31%). In addition, except for SAM_1point, the remaining 8 models show relatively excellent performance in the lung segmentation task, which can effectively meet the segmentation requirements in terms of segmentation results. However, due to the zero_shot property of SAM, for point prompt, SAM cannot judge the target region to be segmented well. At this time, SAM may only take a part of the lung or the complete human body structure as the segmentation target, which leads to poor segmentation results. When the input prompt type is a box, SAM is able to confirm the target region for segmentation more than when the input is point prompt and, therefore, achieves results similar to those of the traditional training model.

We further evaluate the performance of each model on different datasets. To increase the generalizability as well as the reliability of the experimental results, we validated two datasets representing small organs (liver) and lesions

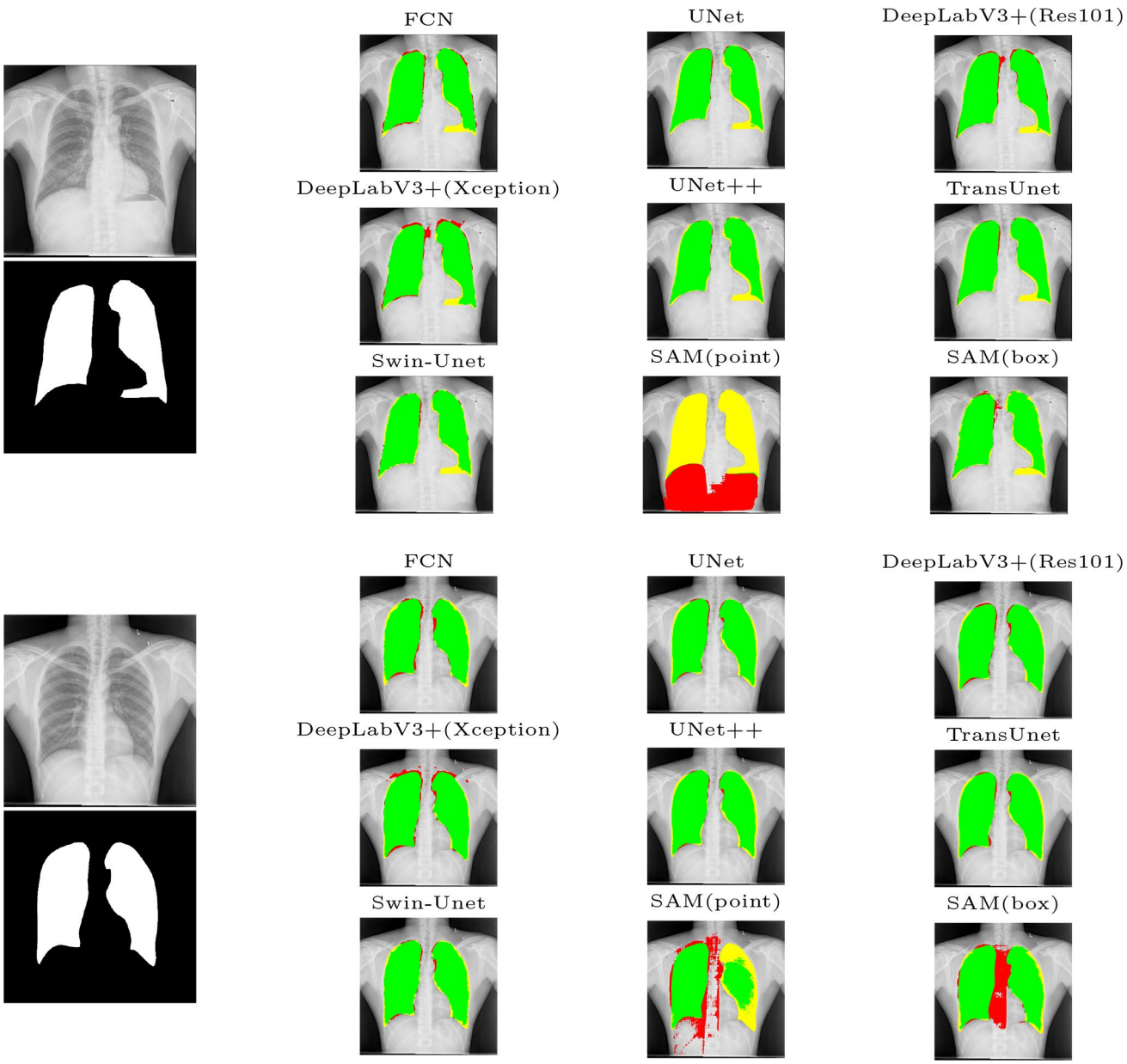

Fig. 10 Tuberculosis Chest X-rays (Shenzhen)dataset segmentation coloring results

(ovarian tumor) in addition to the lung dataset representing larger organs. Tables 4 and 5 list the experimental results for each model, and Figs. 11 and 12 visualize the performance of the models, respectively. Figures 13 and 14 visualize the coloring result where the green area represents tp, the red area represents fp and the yellow area represents fn. Combining the above three datasets, we find that DeepLab V3+ (Resnet101) excels in several evaluation metrics among FCN variants, U-Net, DeepLab variants, and UNet++ using only convolutional structures. It suggests that DeepLab V3+, which synthesizes the DeepLab and the encoder-decoder architecture, has its unique advantages. Surprisingly, using Resnet101 as the backbone in DeepLab V3+ structure to extract the feature information of the image has better results compared to using Xception in the above datasets. After observing the details of the model's segmentation of the boundaries, we found that in the lung segmentation task, the segmentation boundaries of DeepLab series were rougher, while the segmentation results of SAM contained more non-targeted regions. In contrast, in the segmentation task of small organs and diseased tissues, TransUnet and Swin-Unet are better able to obtain edge features of irregular targets.

Additionally, adding the transformer structure to the Encoder-Decoder structure, as done by TransUNet and

Table 4 Ovarian Tumors dataset test index

	DSC↑	HD95↓	IoU↑	Acc↑	Precision↑	Recall↑
FCN	81.54%	26.67	72.71%	96.72%	83.06%	85.18%
U-Net	79.12%	27.63	70.07%	96.88%	80.93%	82.21%
DeepLab V3+(Resnet101)	83.76%	30.46	74.81%	98.46%	**93.34%**	**96.85%**
DeepLab V3+(Xception)	80.96%	45.87	71.42%	98.09%	77.78%	88.23%
UNet++	79.87%	34.43	71.74%	94.48%	77.43%	87.01%
TransUNet	89.18%	22.35	82.73%	99.02%	92.28%	89.20%
Swin-UNet	83.06%	30.80	73.39%	98.22%	78.98%	91.41%
SAM-1point	37.59%	113.41	30.63%	89.01%	33.47%	56.09%
SAM-box	**91.82%**	**11.47**	**85.67%**	**99.05%**	91.89%	92.38%

Bolded datas represent the optimal value of each model for this indicator

Swin-UNet, greatly improves in the segmentation performance of the model. However, Swin-UNet uses the Swin-Transformer module, and the results of completely discarding the convolution do not look as good as TransUNet, which combines both convolution and Transformer features. Combining the advantages of convolution and Transformer may be a more effective way to improve the performance of the model. In addition, for the SAM model, different prompt types significantly affect the segmentation performance of the model, with boxed prompts being more effective compared to simple labeled points.

Discussions

In our experiments, there are a significant number of models employed the U-shaped architecture, and the introduction of the U-Net model was of significant importance in medical image segmentation. U-Net combines an encoder and a decoder, allowing for accurate segmentation by utilizing information at different scales while preserving high-resolution features. This design enables U-Net to achieve more accurate results in medical image segmentation tasks, improving the localization and segmentation precision of lesions. Additionally, U-Net exhibits good scalability and can be improved by adding or adjusting network layers, modifying the network structure, and more. This flexibility enables the U-Net model to be applied to various medical image segmentation tasks and integrated with and optimized alongside other deep learning models.

Initially, Transformer models are not considered promising for medical image segmentation due to their inherent lack of localization ability. However, TransUNet introduces a structure that combines Transformers with convolutional neural networks, forming an effective encoder and improving segmentation performance. On the other hand, if convolutional networks are not combined with Transformers, the final results are not ideal, as demonstrated by Swin-Unet.

SAM has powerful image segmentation capabilities, but its segmentation performance is greatly affected by the type of prompt provided. Some degree of transfer learning is required to fine-tune the model parameters if one wants to use the point-type prompt to get the expected results in a new dataset. However, using box-type prompt has results that match or even surpass other models without the need for fine-tuning. We believe that the box prompt improves the model segmentation effect better than the point prompt.

Challenges and Issues

After several years of development, the problems of deep learning in medical image segmentation applications have come to the fore, and we summarize some of the important open problems and the corresponding effective solutions.

Table 5 Clinical Liver dataset test index

	DSC↑	HD95↓	IoU↑	Acc↑	Precision↑	Recall↑
FCN	79.31%	66.81	71.43%	95.09%	86.79%	78.50%
U-Net	71.57%	77.12	64.22%	86.64%	79.60%	72.04%
DeepLab V3+(Resnet101)	92.74%	61.92	87.22%	98.99%	90.45%	96.05%
DeepLab V3+(Xception)	82.64%	95.95	73.59%	97.48%	82.86%	85.69%
UNet++	83.97%	49.17	78.11%	92.56%	89.48%	81.67%
TransUNet	**96.75%**	**12.28**	**93.94%**	**99.53%**	**96.97%**	96.80%
Swin-UNet	90.76%	27.53	85.37%	97.85%	91.95%	90.59%
SAM-1point	37.59%	113.41	30.63%	89.01%	33.47%	56.09%
SAM-box	93.21%	19.59%	87.79%	98.86%	89.17%	**98.01%**

Bolded datas represent the optimal value of each model for this indicator

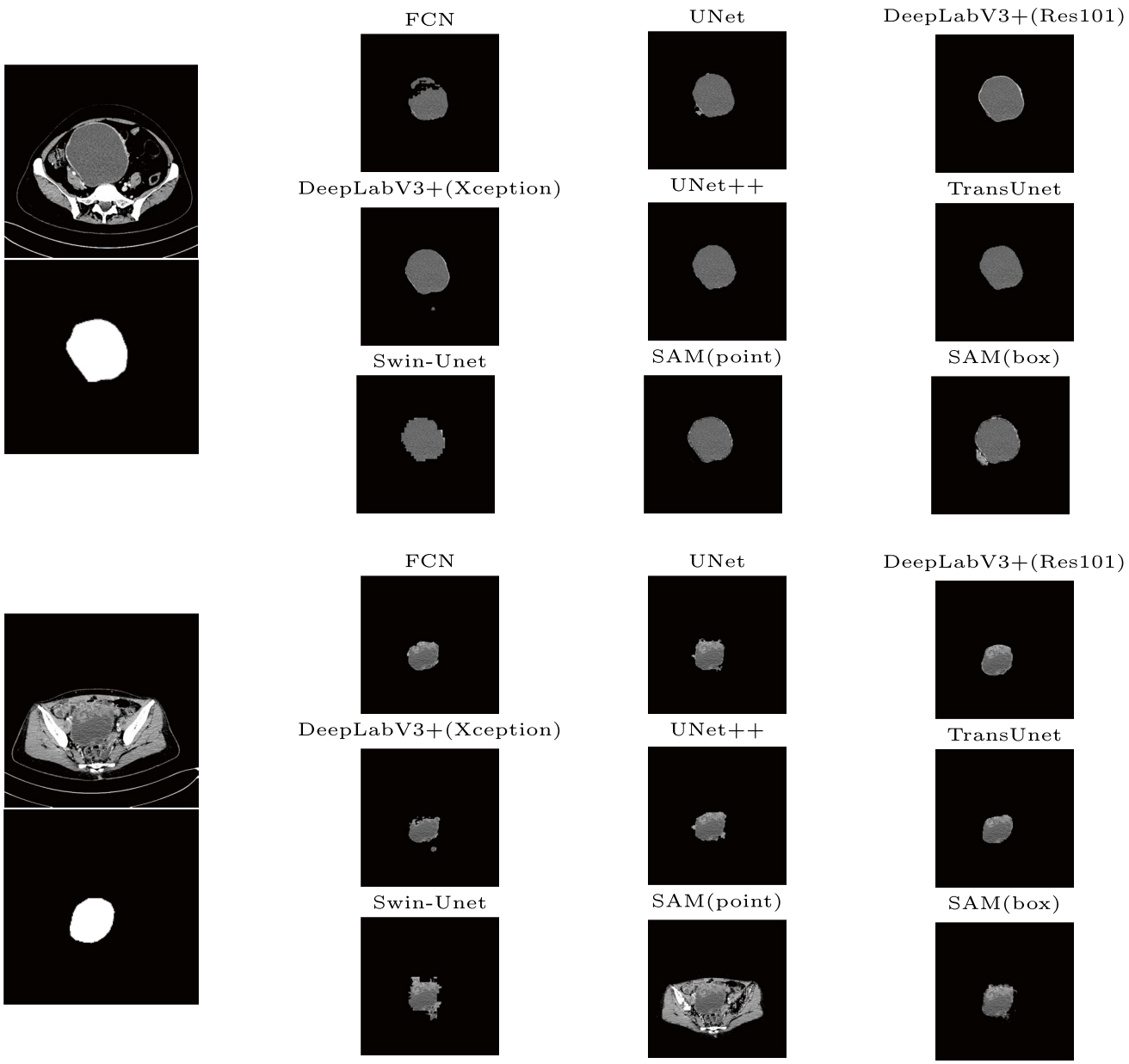

Fig. 11 Ovarian Tumors dataset segmentation results

Medical image labeling is one of the main challenges in applying supervised learning to medical image segmentation. Supervised learning requires the input of a large number of annotated samples to obtain good performance and stable generalization. However, collecting such a large dataset of annotated cases is usually a very daunting task. Medical images require the interpretation of professional clinicians to collect, label, and annotate medical images. Thus the sample sizes of the benchmark datasets currently available are all much smaller than other kinds of [natural] image recognition and segmentation tasks.

Several methods have been widely used to solve this problem. Data augmentation can increase the number of training datasets by applying a set of affine transformations to the samples, such as flipping, rotating, mirroring [33], and enhancing the color (gray) values [34]. Migrating learning from successful models implemented in the same or other domains is another solution to the above problem. Compared to data enhancement, migration learning is a more specific solution that requires only modest computational resources and a smaller amount of labeled data to significantly reduce the error rate in medical image segmentation [35].

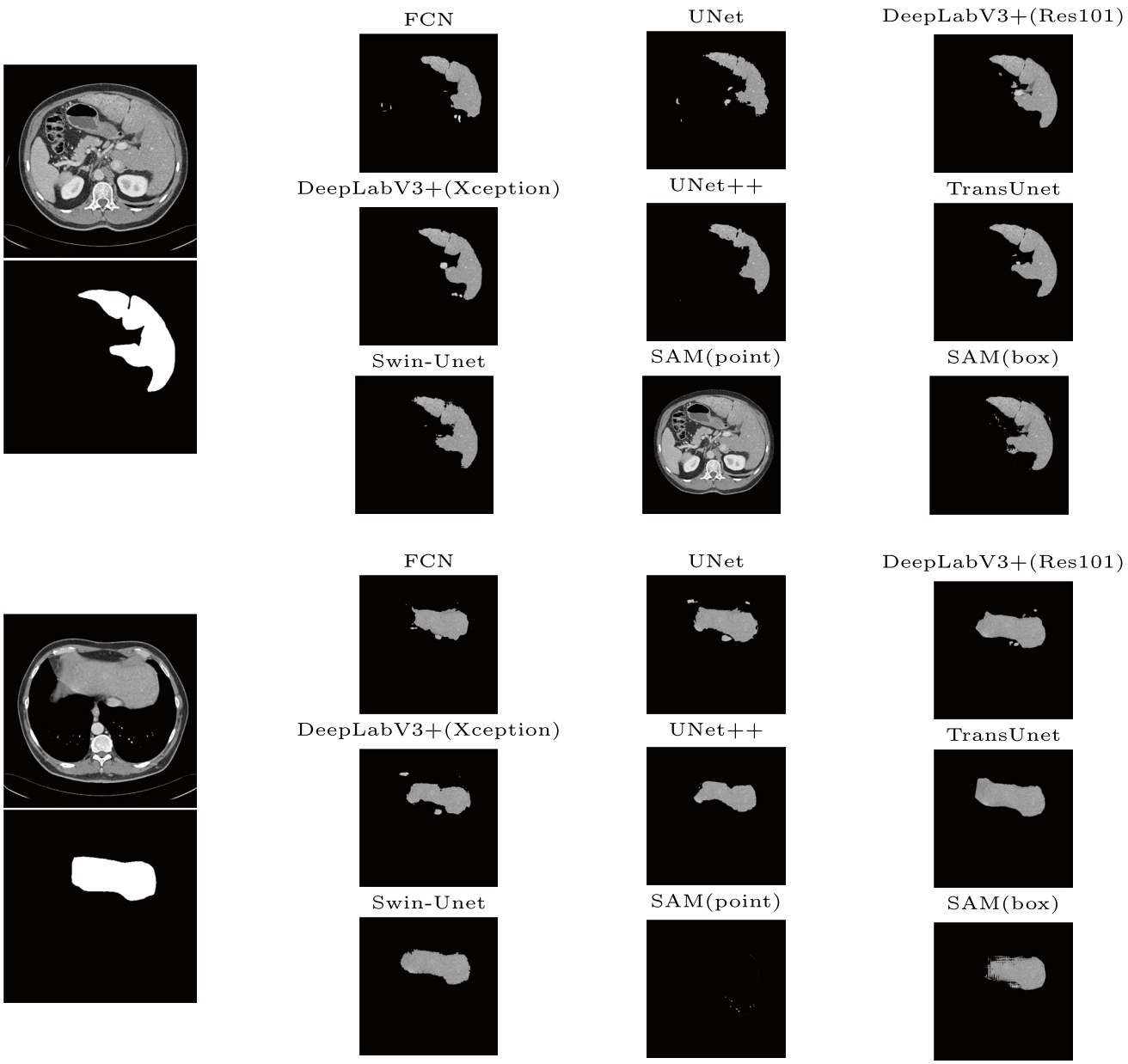

Fig. 12 Clinical Liver dataset segmentation results

Another problem is class imbalance. In medical images of some organs, the region of interest usually occupies a very small portion of the image. Training a network with such data often results in the network focusing on the background. In such cases, it is common to re-weight the training samples or use the attention mechanism, which makes the region of interest have a higher weight. Choosing the right loss function such as the dice coefficient optimizes the network to focus more on small targets.

The heterogeneous appearance of target organs is a major challenge that distinguishes medical image segmentation from natural image segmentation. Organs as well as lesion tissues often differ from patient to patient, and the adhesion produced by lesions will make the boundary of the region of interest blurred, which puts higher requirements on the segmentation performance of the model. Multi-modality-based methods can increase the accuracy of such segmentation [36].

In addition, there are some challenges in model training. One of the biggest challenges in deep learning is overfitting. This occurs when the model remembers the training data too closely and does not generalize well to the test set and other new data, i.e., the model performs mediocrely in

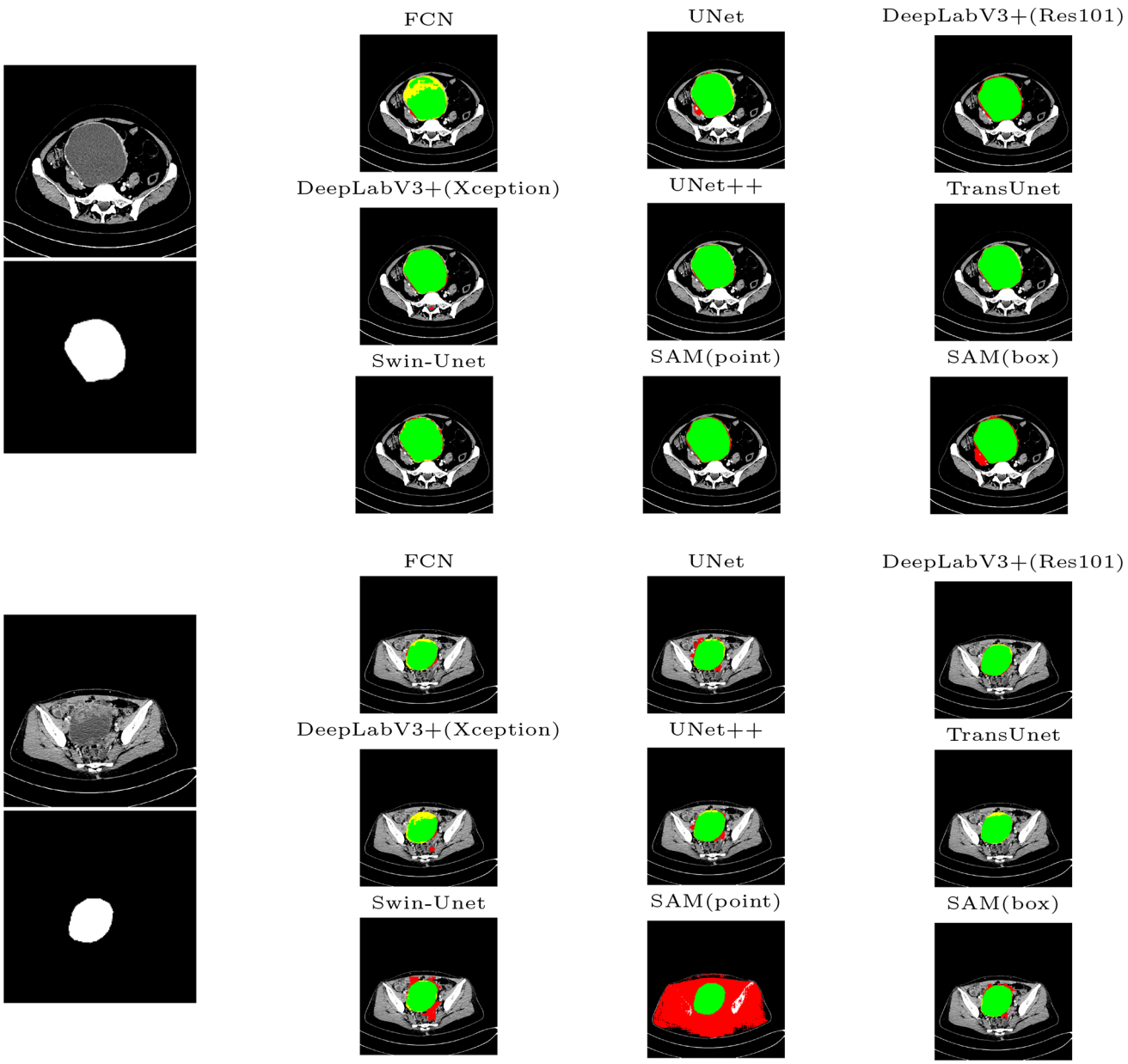

Fig. 13 Ovarian Tumors dataset segmentation coloring results

predicting unknown samples and has poor generalization. Common causes include overlearning, unbalanced sample features, incorrect dataset construction, too many model parameters, too much complexity, and too many iterations of weight learning (Overtraining). Solutions to overfitting include training with more data, performing cross-validation, using dropout or other regularization methods, or using model integration. Gradient vanishing is when the number of layers of a neural network increases, and the weights between the layers that are closer to the input layer cannot be corrected efficiently, resulting in a poorly performing neural network.

Gradient explosion causes the gradient values between network layers to be greater than 1.0, which leads to exponential growth of the gradient, dramatic updating of the network weights, and consequent destabilization of the network. Ways to solve the problem of vanishing, exploding gradients include setting a gradient threshold, using batch normalization, choosing an activation function such as ReLU, and using, for example, residual cross-layer connection structures.

Finally, with the proposal of SAM, the medical segmentation base model became a very desirable model. However, in utilizing SAM, we found that it does perform very well on some datasets, such as ovarian tumor segmentation. But

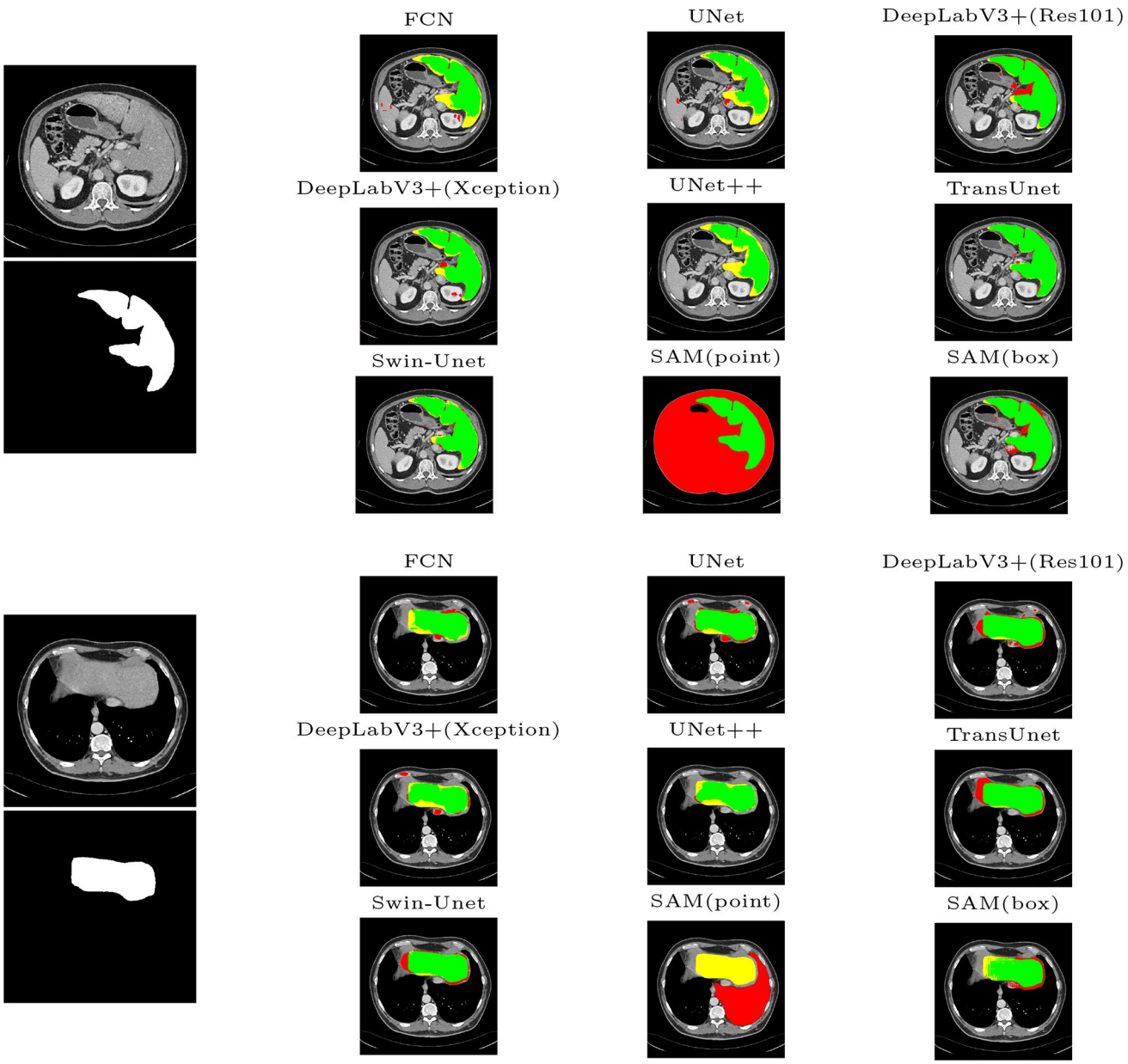

Fig. 14 Clinical Liver dataset segmentation coloring results

the performance is not good enough in some scenarios. It is found that the type or setting of the prompt has a large impact on the experimental performance. How to design a better prompt will make the future research hotspot in this field.

Conclusions

In this paper, we first introduce the general knowledge of medical image segmentation and then review the seven most representative medical image segmentation models, such as FCN, U-Net, DeepLab, UNet++, TransUNet, Swin-Unet, and SAM. We summarize the structural details, as well as the advantages and disadvantages of each model. In addition, we evaluate the quantitative performance of these models on three benchmark datasets, visually demonstrating and analyzing the performance of each model on datasets representing different human structures. Finally, we summarize the existing challenges and propose effective solutions to address the various challenges. Also, to help researchers in related fields understand these models quickly and to model new segmentation tasks, we share all the experimental source code on GitHub and the detailed parameters of the model setup. In the future, we will continue to investigate the application of Foundation Models to medical image segmentation tasks.

Author Contributions [Wenjian Yao, and Mengjuan Liu] contributed to the study conception and design. Material preparation, data collection, and analysis were performed by [Wenjian Yao, Yao Xie, Wei Liao, and Yuheng Chen]. The first draft of the manuscript was written by [Wenjian Yao, Jiajun Bai, and Mengjuan Liu] and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by the Open Project of Network and Data Security Key Laboratory of Sichuan Province (NSD2021-6), Clinical Research and Transformation Fund of Sichuan Provincial People's Hospital (2021LY24), and the Key Research Project of Science and Technology of Sichuan Province(2022YFS0087, 2023YFS0039).

Declarations

Ethics Approval This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of Sichuan Provincial People's Hospital.

Consent to Participate Informed consent was obtained from all individual participants included in the study.

Consent for Publication The authors affirm that patients signed informed consent regarding publishing their data and photographs. **Tuberculosis Chest X-rays dataset** is from the publicly available dataset: Tuberculosis Chest X-rays dataset. **Clinical Liver CT dataset** is from the publicly available dataset: Clinical Liver CT dataset. **Ovarian Tumors dataset**, we obtained all the informed consent. Also, the patient's abdominal images were anonymized so that the images would not identify a patient.

Competing Interests The authors declare no competing interests.

References

1. Cheng, J.Z., Ni, D., Chou, Y.H., Qin, J., Tiu, C.M., Chang, Y.C., Huang, C.S., Shen, D., Chen, C.M.: Computer-aided diagnosis with deep learning architecture: Applications to breast lesions in US images and pulmonary nodules in CT scans. Scientific Reports **6**, 24454 (2016)
2. Golan, R., Jacob, C., Denzinger, J.: Lung nodule detection in ct images using deep convolutional neural networks. In: International Joint Conference on Neural Networks (2016)
3. Christ, P.F., Ettlinger, F., Grün, F., Elshaera, M.E.A., Lipkova, J., Schlecht, S., Ahmaddy, F., Tatavarty, S., Bickel, M., Bilic, P.: Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks (2017)
4. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics **9**(1), 62–66 (1979)
5. Magnier, Baptiste: Edge detection: a review of dissimilarity evaluations and a proposed normalized measure. Multimedia Tools & Applications (2017)
6. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18, pp. 234–241 (2015). Springer
7. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(12), 2481–2495 (2017)
8. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected CRFS. arXiv preprint arXiv:1412.7062 (2014)
9. 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)
10. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision, pp. 205–218 (2022). Springer
11. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision, pp. 213–229 (2020). Springer
12. 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)
13. 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)
14. Aljuaid, A., Anwar, M.: Survey of supervised learning for medical image processing. SN Computer Science **3**(4), 292 (2022)
15. Abdou, M.A.: Literature review: Efficient deep neural networks techniques for medical image analysis. Neural Computing and Applications **34**(8), 5791–5812 (2022)
16. Asgari Taghanaki, S., Abhishek, K., Cohen, J.P., Cohen-Adad, J., Hamarneh, G.: Deep semantic segmentation of natural and medical images: a review. Artificial Intelligence Review **54**, 137–178 (2021)
17. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, pp. 3–11 (2018). Springer
18. Ker, J., Wang, L., Rao, J., Lim, T.: Deep learning applications in medical image analysis. Ieee Access **6**, 9375–9389 (2017)
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
20. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
21. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
22. Chen, L.-C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
23. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)
24. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
25. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence **37**(9), 1904–1916 (2015)
26. Li, C., Tan, Y., Chen, W., Luo, X., Gao, Y., Jia, X., Wang, Z.: Attention unet++: A nested attention-aware u-net for liver ct

image segmentation. In: 2020 IEEE International Conference on Image Processing (ICIP), pp. 345–349 (2020). IEEE
27. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
28. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009 (2022)
29. 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 (2021). PMLR
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
31. Jaeger, S., Candemir, S., Antani, S., Wáng, Y.-X.J., Lu, P.-X., Thoma, G.: Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. Quantitative imaging in medicine and surgery **4**(6), 475 (2014)
32. 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)
33. Milletari, F., Ahmadi, S.-A., Kroll, C., Plate, A., Rozanski, V., Maiostre, J., Levin, J., Dietrich, O., Ertl-Wagner, B., Bötzel, K., *et al*: Hough-CNN: Deep learning for segmentation of deep brain regions in MRI and ultrasound. Computer Vision and Image Understanding **164**, 92–102 (2017)
34. Golan, R., Jacob, C., Denzinger, J.: Lung nodule detection in ct images using deep convolutional neural networks. In: 2016 International Joint Conference on Neural Networks (IJCNN), pp. 243–250 (2016). IEEE
35. Beevi, K.S., Nair, M.S., Bindu, G.: Automatic mitosis detection in breast histopathology images using convolutional neural network based deep transfer learning. Biocybernetics and Biomedical Engineering **39**(1), 214–223 (2019)
36. Urban, G., Bendszus, M., Hamprecht, F., Kleesiek, J., et al.: Multimodal brain tumor segmentation using deep convolutional neural networks. MICCAI BraTS (brain tumor segmentation) challenge. Proceedings, winning contribution, 31–35 (2014)