Towards a Comprehensive, Efficient and Promptable Anatomic Structure Segmentation Model Using 3D Whole-Body CT Scans

Heng Guo[1,2], Jianfeng Zhang[1,2], Jiaxing Huang[1], Tony C. W. Mok[1,2], Dazhou Guo[1], Ke Yan[1,2], Le Lu[1], Dakai Jin[1], Minfeng Xu[1,2]

[1]DAMO Academy, Alibaba Group
[2]Hupan Lab, 310023, Hangzhou, China
gh205191@alibaba-inc.com

Abstract

Segment anything model (SAM) demonstrates strong generalization ability on natural image segmentation. However, its direct adaptation in medical image segmentation tasks shows significant performance drops. It also requires an excessive number of prompt points to obtain a reasonable accuracy. Although quite a few studies explore adapting SAM into medical image volumes, the efficiency of 2D adaptation methods is unsatisfactory and 3D adaptation methods are only capable of segmenting specific organs/tumors. In this work, we propose a comprehensive and scalable 3D SAM model for whole-body CT segmentation, named CT-SAM3D. Instead of adapting SAM, we propose a 3D promptable segmentation model using a (nearly) fully labeled CT dataset. To train CT-SAM3D effectively, ensuring the model's accurate responses to higher-dimensional spatial prompts is crucial, and 3D patch-wise training is required due to GPU memory constraints. Therefore, we propose two key technical developments: 1) a progressively and spatially aligned prompt encoding method to effectively encode click prompts in local 3D space; and 2) a cross-patch prompt scheme to capture more 3D spatial context, which is beneficial for reducing the editing workloads when interactively prompting on large organs. CT-SAM3D is trained using a curated dataset of 1204 CT scans containing 107 whole-body anatomies and extensively validated using five datasets, achieving significantly better results against all previous SAM-derived models.

Code/Data —
https://github.com/alibaba-damo-academy/ct-sam3d

Introduction

Image segmentation is a fundamental task in medical image analysis, with ubiquitous clinical applications such as disease quantification (Iyer et al. 2016; Ferré et al. 2019), computer-aided diagnosis (Roth et al. 2015; Chilamkurthy et al. 2018; Mitani et al. 2020; McKinney et al. 2020), and radiotherapy planning (Jin et al. 2021; Ye et al. 2022; Jin et al. 2022). Despite significant improvements achieved by automatic segmentation methods over the past decade (Wachinger et al. 2018; Isensee et al. 2021; Guo et al. 2024), it remains challenging in daily clinical use due to large variations in medical images, including different imaging protocols, imaging noises/artifacts, and abnormalities or pathological changes among patients (Hesamian et al. 2019; AlBadawy, Saha, and Mazurowski 2018). Interactive segmentation or intelligent image editing with human-in-the-loop techniques (Maleike et al. 2009; Zhao and Xie 2013; Sakinis et al. 2019; Wang et al. 2018, 2019; Ji et al. 2019; Zhang et al. 2021; Koohbanani et al. 2020) are still needed to further refine the segmentation results. Recently, segment anything model (SAM) (Kirillov et al. 2023) shows great success for general-purpose promptable object segmentation in natural images with strong generalization ability and efficient human interaction. Direct deployment of SAM to medical imaging domain exhibits significant performance drops (Wald et al. 2023; Maciej et al. 2023; He et al. 2023; Huang et al. 2024; Deng et al. 2023), but its core design principles of class-agnostic segmentation, prompt encoding, and iterative training scheme can be further exploited and applied to boost the efficiency and accuracy of interactive medical image segmentation.

A few recent studies attempt to fine-tune SAM by incorporating lightweight plug-and-play adapters (Cheng et al. 2023; Chen et al. 2024; Zhang and Liu 2023; Yue et al. 2023; Wu et al. 2023). Simple 2D adaptation methods that completely ignore the intrinsic 3D information require numerous clicks when segmenting hundreds of 2D CT slices, rendering them inapplicable in real clinical practice. In contrast, other

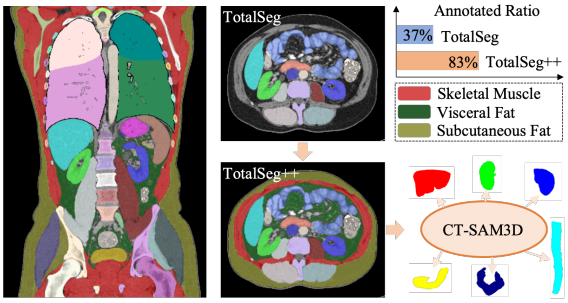

Figure 1: Illustration of the enhanced TotalSeg++ dataset and the versatile 3D promptable CT-SAM3D model. TotalSeg++ complements TotalSeg dataset with added skeletal muscle, visceral and subcutaneous fat annotations.

efforts focus on 3D adaptation by integrating a set of 3D adapters into the SAM architecture. However, these methods have primarily reported segmentation for a limited number of organs/tumors, and their generalizability to a larger set of 3D anatomical categories has not been validated.

In this work, we aim to develop a 3D promptable segmentation model that can interactively segment nearly all anatomic structures within whole-body CT scans with high accuracy and efficiency. To achieve this, we develop a comprehensive, efficient and 3D promptable network, named CT-SAM3D. First of all, we identify several key technical challenges in developing the CT-SAM3D model from scratch. 1) SAM's densely annotated dataset (SA-1B) guarantees that each pixel position in its input space has the opportunity to be positively prompted. It is ideal to have an analogous fully labeled whole-body CT dataset, i.e., each voxel in the valid body region has an anatomic label. Otherwise, some anatomical regions would remain as background, thus not being learned or prompted during training, which limits the model's capability in zero-shot or interactive segmentation scenarios. 2) SAM encodes 2D spatial prompts (points/boxes) by using the sum of one-dimensional random Fourier features (Rahimi and Recht 2007; Tancik et al. 2020) and learned attribute embeddings (positive/negative). However, in full 3D space, this prompt encoding method proves less effective than in 2D. 3) Model's complexity and input data scale can increase dramatically in 3D.

To solve these challenges, we first enrich our whole-body CT scan dataset based on TotalSeg (Wasserthal et al. 2023) by curating the segmentation masks of three important yet under-explored anatomic structures of skeletal muscle, visceral fat, and subcutaneous fat, as illustrated in Fig. 1. This results in a more comprehensive whole-body CT dataset, namely TotalSeg++, where overall $\sim$83% of voxels within the body region are semantically labeled, substantially increased from the previous 37% in TotalSeg dataset. Then, we propose a progressively and spatially aligned prompt encoding method to ensure the model responds to the 3D spatial prompts accurately. Lastly, 3D patch-wise training is necessary to effectively train a 3D SAM model. Yet, if simply training on 3D local image patches, the inference efficiency would be reduced and drastically more clicks are needed to capture the whole spatial context when segmenting organs larger than the local patch dimensions. Therefore, we propose a cross-patch prompt scheme to alleviate this problem.

We outline our main contributions as follows:

- We present a versatile CT-SAM3D model on whole-body CT scans. It is able to segment hundreds of anatomies and achieves new state-of-the-art interactive segmentation results on various datasets.
- We propose two technical novelties: 1) a progressively and spatially aligned prompt encoding method to effectively encode click prompts in local 3D space; 2) a cross-patch prompt scheme to make the local click take effect in a broader spatial context.
- We develop an interactive 3D visualization and segmentation tool with the direct GPU access for model inference. This reports efficient quasi-real-time 3D interactive segmentation performance for the first time.
- We enhance the TotalSeg dataset by adding annotations of three important anatomical structures, resulting in a more comprehensively labeled whole-body CT dataset and facilitating the future research in this field.

Related Work

Segmentation foundation models. The emergence of foundation models in natural language processing (Devlin et al. 2018; Brown et al. 2020) has fostered the development of Vision Foundation Models (VFMs) (Caron et al. 2021; Radford et al. 2021; Oquab et al. 2023; Ramesh et al. 2022). Based on transformer architectures and training on large datasets, VFMs have the potential to enhance various downstream tasks and demonstrate strong zero-shot capabilities. SAM (Kirillov et al. 2023) is the first foundation model for generalized image segmentation, validating its zero-shot capability by segmenting objects in the wild. SegGPT (Wang et al. 2023b) introduces a general interface compatible with various segmentation tasks. SEEM (Zou et al. 2023) offers a unified method using varied prompts to segment and identify objects in images all at once. These methods play an important role in inspiring subsequent works.

SAM adaptation in medical imaging. Substantial disparities exist between natural images and medical images (Shin et al. 2016). The performance of directly applying SAM to medical imaging varies significantly across different objects, anatomies, and modalities (Huang et al. 2024). Considerable efforts have been invested to harness the full potential of SAM in medical imaging. MedSAM (Ma et al. 2024) curates a medical image cohort of $200K$ masks and adapts SAM to medical image segmentation. SAM-Med2D (Cheng et al. 2023) and SAMed (Zhang and Liu 2023) use 2D adapters for medical images. MA-SAM (Chen et al. 2024) and 3DSAM-adapter (Gong et al. 2024) incorporate a set of 3D adapters into each transformer block of the encoder to extract 3D information in medical scans. However, adaptation methods are often restricted to a limited number of organs/tumors. Alternatively, training 3D models from scratch can directly capture 3D contexts. SAM-Med3D (Wang et al. 2023a) simply reforms the 2D SAM into its 3D counterpart to train a 3D SAM model from scratch using $21K$ medical images and $131K$ masks. SegVol (Du et al. 2024) incorporates a zoom-out-zoom-in mechanism into 3D SAM development based on $6K$ CT scans and $150K$ masks. Despite the significant increase in images and masks, the challenges associated with developing a 3D SAM model remain unsolved.

Methodology

CT-SAM3D Architecture

Recall that SAM's ViT (Dosovitskiy et al. 2020) backbone employs a convolution kernel (16×16) to patchify the input image (1024×1024), producing a sequence of 4096 tokens. When dealing with 3D space, we need to avoid the exponential increase in tokens. In terms of feature extraction, it has been extensively shown that hierarchical multi-scale features play a crucial role in semantic segmentation (Ronneberger, Fischer, and Brox 2015; Çiçek et al. 2016; Isensee

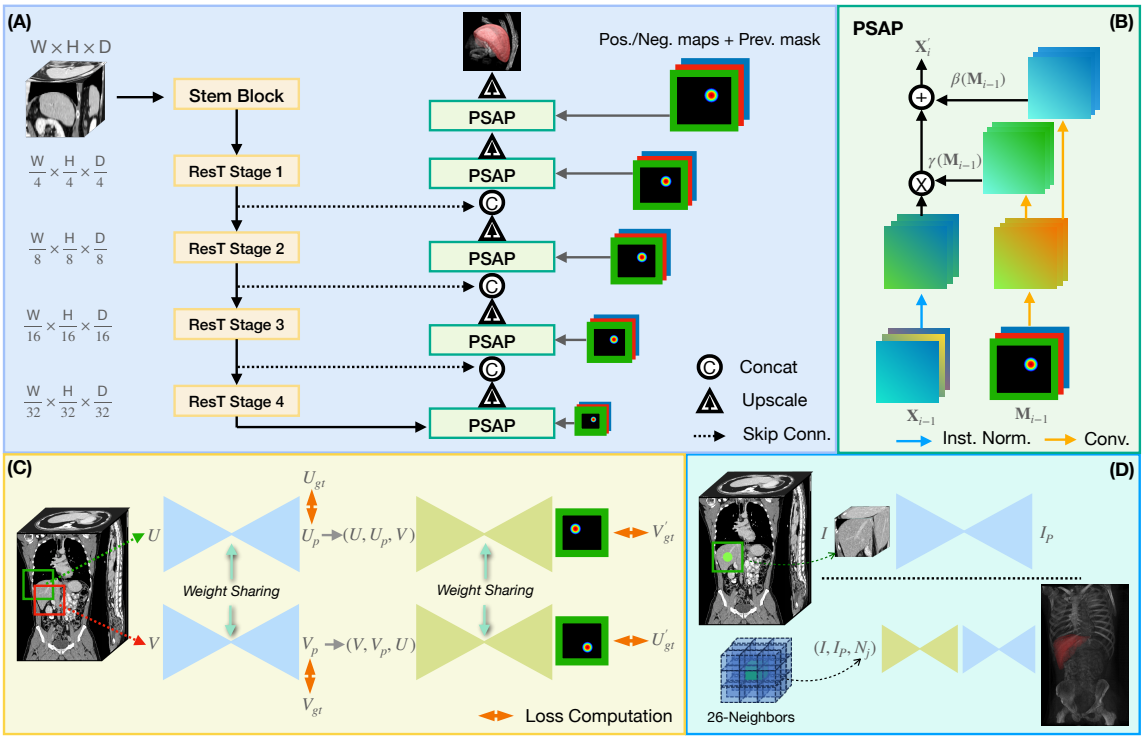

Figure 2: (A) Framework of CT-SAM3D. (B) Details of progressively and spatially aligned prompt. (C) Cross-patch prompt training scheme. (D) Inference on large organs via cross-patch prompting on $\mathbf{N}_j (j \in [1, 26])$, which are the nearest neighbors around the selected patch.

et al. 2021). A recent SAM-derived study reports that simply dividing the ViT into four stages and establishing connections between the feature maps of each stage and the corresponding decoder layers does not confer any advantages (Chen et al. 2024). Taking these into account, we incorporate a hierarchical and memory-efficient Transformer network ResT (Zhang and Yang 2022) and construct skip connections (Ronneberger, Fischer, and Brox 2015) to form another U-shape architecture, as illustrated in Fig. 2(A). The four-stage ResT-based image encoder has a stem block consisting of two consecutive 3D convolution layers as the initial feature extractor. For a given 3D input patch $I \in \mathbb{R}^{W \times H \times D}$, it will be down-scaled by factors of [4, 8, 16, 32] in the hierarchical encoding path. The decoding feature integration is achieved by channel-wise concatenation. The 3D transposed convolutions with kernel size of 2 and stride of 2 are used for feature upscaling. The prompt signals are applied to every decoder stage of the network.

Progressively and Spatially Aligned Prompt

The vanilla SAM's spatial prompt has showcased a robust capability of encoding 2D positions through the utilization of random Fourier features (RFF) (Rahimi and Recht 2007; Tancik et al. 2020). Besides learning to predict masks based on these one-dimensional embeddings, the SAM model also needs to learn an additional embedding to distinguish between positive and negative points. Given a prompted point, it will firstly be normalized to a vector $\mathbf{v} \in [-1, 1]^d$ relative to the input image size, where $d = 2$ in SAM's space. Subsequently, a set of sinusoids is generated as:

$$\begin{aligned} RFF(\mathbf{v}) = [&\cos\left(2\pi \mathbf{b}_1^{\mathrm{T}} \mathbf{v}\right), \sin\left(2\pi \mathbf{b}_1^{\mathrm{T}} \mathbf{v}\right), \ldots, \\ &\cos\left(2\pi \mathbf{b}_m^{\mathrm{T}} \mathbf{v}\right), \sin\left(2\pi \mathbf{b}_m^{\mathrm{T}} \mathbf{v}\right)]^{\mathrm{T}}, \end{aligned} \tag{1}$$

where m is a configurable feature length, and $\mathbf{b}_j (j \in [1, m])$ is sampled from an isotropic distribution. The final prompt encoding of $\mathbf{v}$ is $PE(\mathbf{v}) = RFF(\mathbf{v}) + \mathbf{e}$, where $\mathbf{e}$ is the learned positive/negative embedding. We experimentally found that simply adapting this technique to 3D is less effective than in 2D, as the spatial alignment between the prompt embedding and the 3D position is more challenging to learn. This may lead to unexpected interactive behavior.

Partially inspired by the spatially-adaptive normalization (Park et al. 2019) that effectively preserves the location geometry of the semantic label map, we propose a progressively and spatially aligned prompt (PSAP) for 3D prompt encoding. In contrast to SAM's prompt encoding, our proposed method encodes positive and negative points into two separate click maps, eliminating the need to learn the positive/negative attribute embedding for prompts. This principle is also adopted by some previous 2D interactive segmentation methods (Xu et al. 2016; Sofiiuk, Petrov, and Konushin 2022; Chen et al. 2022; Liu et al. 2023). Our PSAP differs in that it omits the need for forwarding through a computationally intensive image encoder, ensuring faster interactive responses. Concretely, for a given click point $\mathbf{v} = (x, y, z)$, a Gaussian heatmap is generated around

this point. This heatmap is regarded as feature map $\mathbf{P}$ if it is a positive click; otherwise, it is feature map $\mathbf{N}$. Subsequently, the mask prediction $\mathbf{Y}$ from the previous iteration (filled with zeros in the initial iteration), is concatenated with $\mathbf{P}$ and $\mathbf{N}$ to form the composite feature map $\mathbf{M}$ with the order of $[\mathbf{P}, \mathbf{N}, \mathbf{Y}]$. Then, let $\mathbf{X}_{i-1}$ denote the feature map output of the $(i-1)^{th}$ decoder layer, $\mathbf{M}$ will be downsampled to $\mathbf{M}_{i-1}$ that has the same spatial dimensions as $\mathbf{X}_{i-1}$. As illustrated in Fig. 2(B), we compute $\gamma(\mathbf{M}_{i-1}) = Conv_{\gamma}(Conv(\mathbf{M}_{i-1}))$ and $\beta(\mathbf{M}_{i-1}) = Conv_{\beta}(Conv(\mathbf{M}_{i-1}))$, then the output of the proposed PSAP is formulated as follows:

$$\mathbf{X}_i^{'} = \gamma(\mathbf{M}_{i-1})\mathrm{IN}(\mathbf{X}_{i-1}) + \beta(\mathbf{M}_{i-1}), \tag{2}$$

where $\gamma(\mathbf{M}_{i-1})$ and $\beta(\mathbf{M}_{i-1})$ are the learned modulation parameters of the instance normalization (IN) layer (Ulyanov, Vedaldi, and Lempitsky 2016), and they both have the same spatial dimensions as $\mathbf{X}_{i-1}$. Finally, $\mathbf{X}_i^{'}$ will be upscaled and concatenated with the corresponding encoder features to form $\mathbf{X}_i$. This technique enables the model to encode 3D clicks along the decoding path progressively and spatially aligned, resulting in improved efficiency and accuracy for 3D interactive segmentation.

Cross-Patch Prompt for Large Anatomy

Due to limitations in GPU memory and computation efficiency, it is not feasible to take as input an entire 3D volume (e.g., $512 \times 512 \times 192$ for a typical Thoracic CT scan) for network training. However, cropping 3D medical images into smaller sub-volumes can result in truncation of larger or tubular-shaped anatomies, e.g., liver or aorta. Training on only 3D local image patches reduces inference efficiency and requires more manual clicks. To alleviate this problem, we introduce an innovative cross-patch prompt (CPP) approach to capture a broader spatial context. Specifically, beyond the promptable segmentation model (denoted as $\mathcal{S}$ and condensed as the light blue modules in Fig. 2(C)), we add a light-weight encoder-decoder sub-network (denoted as $\mathcal{P}$ and condensed as the light green modules in Fig. 2(C)), which is dedicated to predicting prompts for neighboring patches given a clicked patch and its mask prediction.

To achieve this, we sample two patches with overlapping regions in each anatomy during training iterations. As illustrated in Algorithm 1, the training pipeline is executed in N iterations to simulate a real-world interactive scenario. Let U and V denote two patches sharing overlapping regions, U_C and V_C denote the sampling clicks from the simulated user. The outputs of the segmentation network for the two input patches are U_p and V_p, and the associated segmentation loss for two local patches $\mathcal{L}_{local}$ is calculated (line 3), where $\mathcal{L}_{seg}(\cdot)$ denotes the combination of Focal loss (Lin et al. 2017) and Dice loss (Milletari, Navab, and Ahmadi 2016). Then, the CPP prediction module takes as inputs the crossed and concatenated tuples (U, U_p, V) and (V, V_p, U) to form a mutual anatomical region identification task (line 4). In this task, we aim to predict a heatmap that can be used as click prompts as aforementioned. The ground truth heatmaps $V_{gt}^{'}$ and $U_{gt}^{'}$ are constructed based on the centroids

Algorithm 1: Training Algorithm of CT-SAM3D

Require:
U, V: Patch samples with overlapping;
U_{gt}, V_{gt}: Ground truth masks;
U_C, V_C: Sampling clicks;
ϵ: Learning rate;
N: Number of iterations per sample;
$\mathcal{S}, \mathcal{P}$: Segmentation and CPP prediction modules.
Ensure: Optimal network parameters θ

1: **while** $N > 0$ **do**
2: $U_p = \mathcal{S}(U, U_C), V_p = \mathcal{S}(V, V_C)$
3: $\mathcal{L}_{local} = \mathcal{L}_{seg}(U_p, U_{gt}) + \mathcal{L}_{seg}(V_p, V_{gt})$
4: $V_C^{'} = \mathcal{P}(U, U_p, V), U_C^{'} = \mathcal{P}(V, V_p, U)$
5: $\mathcal{L}_{cross} = \mathcal{L}_{cpp}(U_C^{'}, U_{gt}^{'}) + \mathcal{L}_{cpp}(V_C^{'}, V_{gt}^{'})$
6: $\mathcal{L} = \mathcal{L}_{local} + \mathcal{L}_{cross}$
7: $\theta \leftarrow \theta - \epsilon \nabla_{\theta} \mathcal{L}$
8: Update U_C, V_C according to the error regions
9: $N \leftarrow N - 1$
10: **end while**

of the foregrounds in V_{gt} and U_{gt}, respectively. After obtaining the CPP predictions $U_C^{'}$ and $V_C^{'}$, we compute $\mathcal{L}_{cpp}(\cdot)$ using a mean squared error loss, then the cross-patch loss $\mathcal{L}_{cross}$ is calculated (line 5). The CT-SAM3D is then updated (line 7), considering both the local segmentation loss and cross-patch prompt prediction loss. We omit the mask input and its updating in the pseudo-code for simplicity. This training process is also illustrated in Fig. 2(C).

When segmenting a large anatomy such as the liver or aorta, a straightforward approach involves starting with an initial click to obtain the corresponding mask of the local patch. Then, the number of clicks can be gradually increased in the uncovered areas until the complete result is achieved. Yet, this is not user-friendly. As shown in Fig. 2(D), CPP can reduce workloads by utilizing the 26 nearest spatial neighboring patches surrounding the selected patch, thereby attaining precise segmentation results with fewer clicks.

Experiments

We first introduce the datasets used in our experiments.

TotalSeg dataset consists of 1204 CT scans with 104 anatomical structures annotated (Wasserthal et al. 2023). Using an in-house developed muscle/fat segmentation model with manual examination and curation, we enhance the TotalSeg dataset by introducing annotations of three anatomies, i.e., skeletal muscle, visceral fat, and subcutaneous fat. This results in a comprehensive whole-body CT dataset where overall ∼83% of voxels within the body possesses a semantic label. We refer to this enhanced dataset as **TotalSeg++**. Details of the data curation process are in our supplementary material. We follow the original data split, using 1139 CT scans for training and 65 for internal testing.

FLARE22 is proposed in an abdominal organ segmentation challenge held at MICCAI 2022 (Ma et al. 2023). The 13 labeled organs include the liver, spleen, pancreas, etc. The offline test set of 20 CT cases is used for external testing.

BTCV is also an abdominal challenge dataset (Landman et al. 2015) that includes 30 CT scans with annotations for 13 organs, differing slightly from FLARE22 as it lacks duodenum but includes portal vein and splenic vein annotations. The total 30 CT scans are also used for external testing.

MSD-Pancreas and **MSD-Colon** are two tumor segmentation datasets from Medical Segmentation Decathlon challenge (Antonelli et al. 2022). They contain 281 and 126 abdominal CT scans respectively. They are used to validate the model's zero-shot segmentation capability.

Evaluation Protocol

To maximize the reproducibility, we follow the practice in (Wang et al. 2023a; Kirillov et al. 2023; Sofiiuk, Petrov, and Konushin 2022) to simulate the real-world interactive scenario. Specifically, the first simulated click point is randomly sampled from the foreground region, and the subsequent point is sampled iteratively using the farthest point from the boundary of error regions. We measure 3D organ-wise Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD) (Maier-Hein, Menze et al. 2022) after N point prompts, where $N \in \{1, 3, 5, 7, 9\}$. Hence, predictions from 2D SAM-derived methods need to be heavily merged before calculating the metrics. Note that we take into account the number of prompt points in the corresponding model space for evaluation. In other words, for a 2D SAM-derived model, N indicates the number of prompts used in each 2D CT slice (not multiplied by the number of slices), whereas for a 3D SAM-derived model, N represents the number of prompts in a whole 3D CT scan. Even under this biased evaluation protocol, our experiments demonstrate that CT-SAM3D significantly outperforms all other 2D SAM-derived methods by a large margin of at least 10% DSC, while actually requiring fewer clicks.

Comparing methods. SAM (Kirillov et al. 2023) and recent SAM-inspired medical image segmentation models are primarily and extensively compared, including MedSAM (Ma et al. 2024), SAMed (Zhang and Liu 2023), MA-SAM (Chen et al. 2024), SAM-Med2D (Cheng et al. 2023), SAM-Med3D (Wang et al. 2023a) and SegVol (Du et al. 2024). Note that SAMed and MA-SAM disable the prompt encoding module, so only a fixed number of organs that have appeared in their training datasets can be segmented.

Implementation details. Our implementation is built upon PyTorch (Paszke et al. 2019). CT-SAM3D is trained using AdamW optimizer (Loshchilov and Hutter 2017) with an initial learning rate of $1e{-}4$. The total training process consists of 1000 epochs, with the first 100 epochs to linearly warm up. The learning rate is then reduced by a factor of 10 at the 800th epoch. Our model is trained on 8 A100 GPUs, with a batch size of 4 per GPU and a sampling number of 8 per volume. The number of iterations per batch is set to be 5. The first iteration only uses points as prompts, and the subsequent iterations use both updated points and previous masks as prompts simultaneously. For $\mathcal{L}_{seg}(\cdot)$ in Algorithm 1, we use a combination of Focal loss (Lin et al. 2017) and Dice loss (Milletari, Navab, and Ahmadi 2016) with coefficients of 0.2 and 0.8, respectively. For data preprocessing, we process all data to have an isotropic spacing of 1.5 mm and apply normalization to scale it to the range of [0, 1]. The patch size is set as $(64 \times 64 \times 64)$. Random cropping is employed to obtain training samples. Notably, CT-SAM3D is trained from scratch. To facilitate the developing and validation of CT-SAM3D, we have developed an interactive 3D visualization and segmentation tool with quasi-real-time responses, details of which are provided in the supplementary.

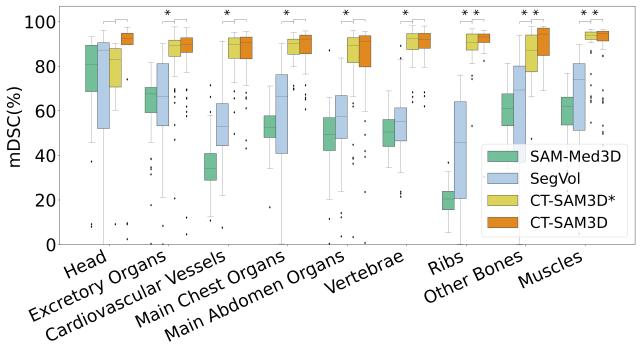

Figure 3: Grouped boxplots. "CT-SAM3D*" (lemon color) denotes degraded results when trained on TotalSeg. "*" above the boxes denotes $p < 0.05$.

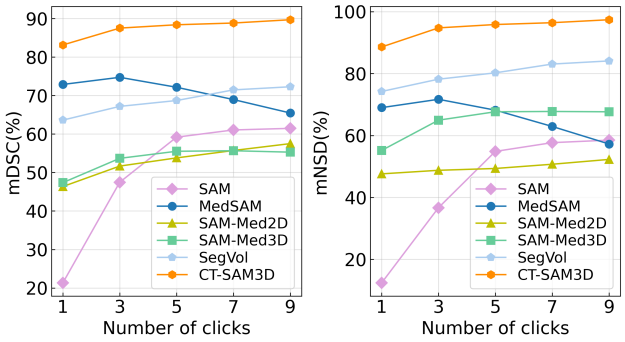

Figure 4: FLARE22 results under increasing clicks.

Main Results

Internal testing on TotalSeg++ is summarized in Fig. 3. We found that only SAM-Med3D and SegVol can achieve meaningful results on over 100 anatomies during testing, partly because TotalSeg is a subset of their training sets. Therefore, only these two methods are taken as baselines in this experiment. We group the annotated anatomies into 9 groups (head, excretory organs, cardiovascular vessels, main chest organs, main abdomen organs, vertebrae, ribs, muscles, and other bones). Compared to SAM-Med3D and SegVol, our CT-SAM3D yields substantially better performance across all groups. These results suggest that simply replacing 2D operations in SAM with its 3D counterparts does not work well, although the TotalSeg dataset is already included in their training set. In contrast, CT-SAM3D uses only 1139 CT scans for training and achieves evidently superior performance than SAM-Med3D and SegVol, by proposing more suitable 3D SAM network architecture with effective prompt

Methods	Liver	KidneyR	Spleen	Pancreas	Aorta	IVC	RAG	LAG	GB	ESO	Stomach	DUO	KidneyL	mDSC↑
SAM	86.0±5.5	87.6±8.7	84.5±8.9	53.4±10.6	77.5±16.1	44.5±16.3	19.4±14.0	33.9±15.0	52.4±16.4	35.2±6.8	68.0±11.2	44.4±12.4	82.6±11.3	59.2
MedSAM	93.0±3.1	90.0±5.3	89.1±11.0	73.5±9.6	82.5±19.7	76.5±19.4	36.0±23.6	48.7±22.6	56.4±27.1	64.7±19.9	84.0±13.0	53.9±11.7	89.7±7.9	72.2
SAMed	92.4±5.1	74.6±28.9	90.6±6.3	65.0±19.6	83.4±11.7	–	–	–	71.3±25.4	–	76.9±21.5	–	77.2±27.4	78.9*
MA-SAM	92.8±5.4	80.0±20.1	87.6±13.4	74.1±14.1	86.4±10.2	80.1±16.7	45.0±16.3	46.9±18.2	77.1±13.2	70.5±17.8	75.9±22.0	–	77.3±25.4	74.5*
SAM-Med2D	91.4±5.8	83.7±17.3	83.9±15.2	58.8±18.7	60.6±22.5	18.6±10.4	10.6±9.7	27.1±12.4	32.9±21.6	28.1±13.3	72.9±16.6	45.4±19.8	86.0±16.8	53.8
SAM-Med3D	85.4±13.2	84.2±9.5	84.7±11.8	46.9±14.3	60.4±10.7	44.5±13.4	32.6±20.9	35.3±18.3	56.0±19.4	32.6±16.4	46.9±19.8	27.4±13.6	84.9±6.9	55.5
SegVol	83.9±25.3	71.7±30.6	75.9±28.8	69.4±16.1	83.1±12.1	80.3±13.9	42.1±13.3	49.7±22.6	55.6±31.1	69.6±8.4	81.1±20.7	55.6±19.8	75.1±22.6	68.7
CT-SAM3D	**95.6**±2.0	**95.0**±1.8	**96.1**±4.4	**83.6**±12.0	**94.5**±2.8	**91.8**±4.7	**78.4**±18.0	**82.5**±4.0	**88.4**±8.1	**82.9**±18.1	**92.3**±4.4	**73.2**±16.8	**94.8**±1.4	**88.4**

Methods	Liver	KidneyR	Spleen	Pancreas	Aorta	IVC	RAG	LAG	GB	ESO	Stomach	DUO	KidneyL	mNSD↑
SAM	61.4±9.6	73.2±18.2	67.8±15.4	57.0±11.1	77.8±20.0	33.6±13.7	40.0±22.3	53.8±15.5	46.2±20.4	42.3±9.6	50.4±13.6	46.8±13.0	62.8±21.4	54.9
MedSAM	74.7±10.4	78.6±17.3	74.0±19.8	72.1±14.9	79.6±23.5	71.8±22.6	53.2±28.1	57.8±24.0	46.5±32.2	67.5±21.3	72.8±18.3	53.4±10.5	84.4±15.9	68.2
SAMed	86.1±9.6	74.6±27.0	89.3±11.6	77.4±17.2	87.5±11.4	–	–	–	81.5±22.5	–	72.9±18.4	–	75.4±25.8	80.6*
MA-SAM	87.6±10.7	80.3±18.1	87.9±14.8	85.9±10.8	89.8±10.2	87.7±13.6	59.6±13.6	62.0±18.9	88.1±13.1	85.7±17.5	71.4±19.7	–	79.6±24.0	80.5*
SAM-Med2D	79.5±18.3	72.9±27.2	72.0±27.2	59.1±22.8	47.6±24.7	15.0±6.7	16.1±11.6	41.5±16.5	24.6±20.2	29.4±15.7	58.8±21.5	43.2±18.8	81.8±25.1	49.3
SAM-Med3D	76.3±19.2	82.8±11.6	84.5±14.3	59.4±14.5	70.3±8.7	62.2±11.0	66.6±28.6	70.0±23.1	76.3±12.8	58.2±17.9	48.8±19.2	40.7±15.2	83.5±10.0	67.7
SegVol	79.7±25.0	74.5±29.4	77.1±26.6	83.3±13.6	91.7±9.1	92.2±10.4	75.6±12.8	76.6±23.2	68.9±33.4	91.1±7.1	84.3±21.6	69.6±19.7	78.4±23.8	80.2
CT-SAM3D	**94.7**±4.9	**98.0**±2.6	**98.1**±7.5	**92.5**±15.4	**98.4**±3.4	**97.4**±4.9	**95.2**±19.3	**99.7**±0.3	**97.6**±6.1	**94.3**±18.5	**96.1**±6.3	**85.3**±16.9	**98.4**±2.3	**95.8**

Table 1: Organ-specific DSC (%) and NSD (%) evaluation on FLARE22. The results were obtained after 5 clicks. Abbreviations: "IVC"-Inferior Vena Cava, "RAG"-Right Adrenal Gland, "LAG"-Left Adrenal Gland, "GB"-Gallbladder, "ESO"-Esophagus, "DUO"-Duodenum, "mDSC"-mean DSC, "mNSD"-mean NSD. (*) denotes that these results gather exclusively valid values.

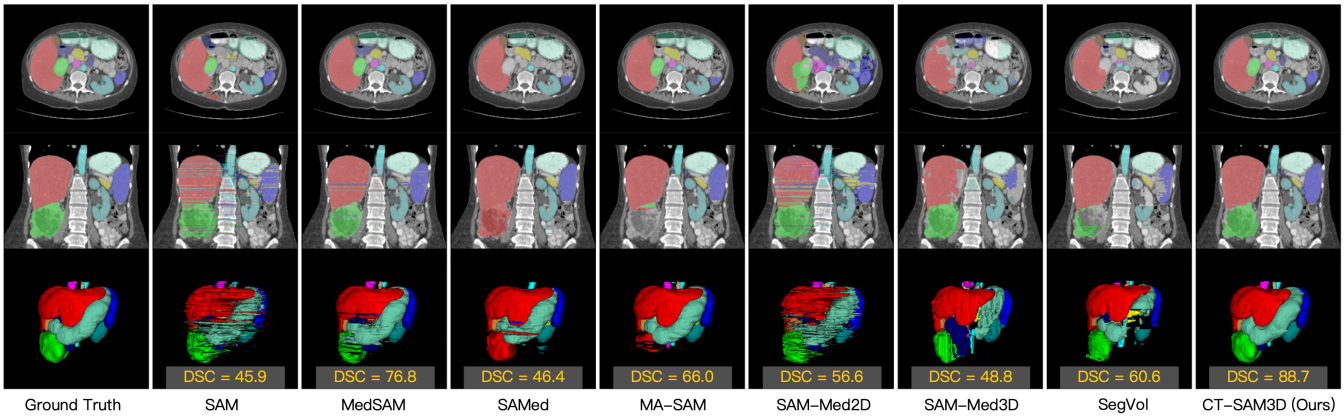

Figure 5: Qualitative results of different methods on a subject who exhibits severe renal pathology (green region). The first row is an axial slice, the second row is a coronal slice, and the last row shows the 3D volume rendering. DSC (%) scores are mentioned for each method.

encoding. To investigate the impact of the newly added annotations, we build a degraded CT-SAM3D* that is trained only on TotalSeg. As observed, the median DSC of most groups in CT-SAM3D surpasses that of CT-SAM3D*, and several groups (ribs, muscles, and other bones) achieve statistically significant improvements ($p < 0.05$), indicating the newly added annotations have a positive impact on our model. Moreover, CT-SAM3D* still performs significantly better than almost all groups in SegVol and SAM-Med3D, indicating the superiority of our model architecture. Detailed results for each anatomy, including the newly added annotations, are provided in the supplementary material.

External testing on FLARE22 is summarized in Table 1 and Fig. 4. Table 1 presents the detailed organ-wise segmentation results when the prompt number $N = 5$, while Fig. 4 illustrates the organ-averaged segmentation accuracy with respect to the prompt number N, as N varies from 1 to 9. As shown in Table 1 and Fig. 4, the proposed CT-SAM3D generalizes well to the external testing and performs significantly better as compared to other SAM-derived models with a large margin of $>\sim$10% DSC and $>\sim$15% NSD. Several interesting observations can be made. 1) The very recent 3D SAM models (SAM-Med3D and SegVol) do not generalize well, although they are trained on $21K$ and $6K$ medical images respectively. 2) Although SAMed and MA-SAM disable the prompt encoding module, they achieve the 2nd and 3rd best accuracy (if not considering the missing targets). This demonstrates the effectiveness of adaptation methods to some degree. However, these adaptation methods target on segmenting only a small fixed number of organs, limiting their general interactive segmentation ability. 3) We see from Fig. 4 that with more point

Methods	1 click		3 clicks	
	mDSC↑	mNSD↑	mDSC↑	mNSD↑
SAM	14.2	6.8	39.3	31.3
MedSAM	64.2	61.8	67.9	67.8
SAM-Med2D	42.3	51.3	48.1	48.3
SAM-Med3D	41.9	51.5	47.5	61.2
SegVol	59.5	71.3	63.4	76.2
CT-SAM3D	**78.4**	**88.4**	**82.2**	**93.9**

Table 2: DSC (%) and NSD (%) results on BTCV dataset.

Methods	Pancreas Tumor		Colon Cancer	
	mDSC↑	mNSD↑	mDSC↑	mNSD↑
nnU-Net	41.65	62.54	43.91	52.52
nnFormer	36.53	53.97	24.28	32.19
Swin UNETR	40.57	60.05	35.21	42.94
UNETR++	37.25	53.59	25.36	30.68
3D UX-Net	34.83	52.56	28.50	32.73
SAM	30.55	32.91	39.14	42.70
3DSAM-adapter	57.47	**79.62**	49.99	**65.67**
CT-SAM3D	**59.60**	77.93	**50.68**	64.14

Table 3: Zero-shot tumor segmentation of CT-SAM3D.

prompts, CT-SAM3D's segmentation accuracy continues to increase, illustrating its interactive segmentation capacity. In contrast, with even 9 prompts per input sample, other methods still obtain undesirable segmentation accuracy, ranging from 55% to 75% DSC or from 50% to 85% NSD. 4) Not all methods improve consistently when the number of clicks increases, e.g., MedSAM starts to degrade when $N > 3$, indicating its deficiency in understanding/encoding user's prompts. A qualitative example is shown in Fig. 5.

External testing on BTCV is illustrated in Table 2. It can be seen that with only 1 click prompt, CT-SAM3D has a mDSC of 78.4% and mNSD of 88.4%, substantially higher than all other interactive SAM-derived models. When increasing the number of clicks to 3, CT-SAM3D improves with 3.8% mDSC (from 78.4% to 82.2%) and 5.5% mNSD (from 88.4% to 93.9%). The obtained performance (3 clicks per organ) is also comparable to the fully supervised segmentation model trained on BTCV. These results demonstrate CT-SAM3D's generalization ability to unseen datasets, indicating the effectiveness for potential real clinical usage.

Zero-shot tumor segmentation. Beyond anatomical structure segmentation tasks, we have applied CT-SAM3D to more challenging endeavors, i.e., tumor segmentation, to investigate its zero-shot capabilities. A previous SAM adaptation method, 3DSAM-adapter (Gong et al. 2024), has also reported results on the aforementioned two tumor datasets. 3DSAM-adapter are fine-tuned on each dataset, followed by testing on a randomly selected 20% of the data. Our significant distinction is that we do not conduct any fine-tuning operation, resulting in a truly zero-shot testing. For a fair comparison, we report results on the same test splits used in 3DSAM-adapter. The results are summarized in Table 3. Some state-of-the-art automatic segmentation methods are also included (Isensee et al. 2021; Zhou et al. 2023; Tang et al. 2022; Shaker et al. 2022; Lee et al. 2022). As observed, a promptable, human-in-the-loop approach can greatly elevate the upper-bound of challenging tumor segmentation results. CT-SAM3D outperforms nnU-Net by 17.95% mDSC on pancreas tumor segmentation, primarily benefiting from input prompts that directly indicate the location of the tumors. When compared to the specifically fine-tuned 3DSAM-adapter, our CT-SAM3D achieves comparable performance on both pancreas tumor segmentation and colon cancer segmentation tasks under 10 clicks, maintaining a slight edge on mDSC yet exhibiting a minor shortfall in mNSD, even under a more challenging zero-shot setting. A qualitative comparison is shown in our supplementary material, where we showcase results of SAM, 3DSAM-adapter, and our proposed CT-SAM3D under gradually increased click prompts. This comparison further demonstrates the excellent zero-shot capability of CT-SAM3D.

I.Enc.	w/ RFF	w/ PSAP	FLARE22		BTCV	
			mDSC↑	mNSD↑	mDSC↑	mNSD↑
ResT	✓		73.1	81.4	61.5	72.8
ResT		✓	**87.5**	**94.8**	**82.2**	**93.9**
UNet		✓	86.4	93.8	79.9	92.0

Table 4: Effectiveness of PSAP. "I.Enc."-image encoder.

Organ	w/ CPP	FLARE22		BTCV	
		mDSC↑	mNSD↑	mDSC↑	mNSD↑
Liver		41.9	31.5	38.1	29.7
Liver	✓	**87.1**	**77.1**	**80.5**	**70.3**
Aorta		51.9	55.7	46.8	51.1
Aorta	✓	**61.6**	**64.1**	**57.7**	**61.9**

Table 5: Effectiveness of CPP on large organs under 1 click.

Ablation Study Results

The effectiveness of PSAP. To find an effective prompt encoding method, we have carefully compared the performance of random Fourier features (RFF) (Rahimi and Recht 2007; Tancik et al. 2020) and our proposed progressively and spatially aligned prompt (PSAP) on the tasks of FLARE22 and BTCV. It is observed that, under the same image encoder (ResT), PSAP achieves improvements of 20.7% and 21.1% in mDSC and mNSD, respectively, compared to RFF on BTCV, as shown in Table 4. Similar observations are obtained on FLARE22. Interestingly, when we replace the image encoder with the UNet (Çiçek et al. 2016) structure, which is more common in medical image analysis, we can still obtain very competitive results. This implies that an effective prompt encoding mechanism is essentially crucial in 3D interactive segmentation. It also reveals that PSAP can

serve as a plug-and-play module to be incorporated into different hierarchical backbone network structures to construct a stronger interactive segmentation model.

The effectiveness of CPP. To investigate the effectiveness of cross-patch prompt (CPP) on segmenting large organs, we conduct a 1-click experiment on liver and aorta segmentation tasks (Table 5). As expected, the results are quite disappointing without CPP due to large organ dimensions. Yet, a single click with enabled CPP mechanism can boost mDSC performance of liver by 45.2% and 42.4% on FLARE22 and BTCV, respectively. Even on a tubular-shaped aorta structure, CPP can bring $\sim$10% gains both in mDSC and mNSD scores. It should also be noted that there is still room for further improvement as the number of clicks increases.

Conclusion

In this work, we present a comprehensive, efficient and 3D promptable model on whole-body CT scans. Instead of adapting SAM, we directly develop a pure 3D promptable model utilizing a more comprehensively labeled CT dataset (i.e., TotalSeg++). To train CT-SAM3D effectively using 3D local image patches, we propose two key technical developments to encode the click prompt in local 3D space and conduct the cross-patch prompt scheme to reduce clicks when segmenting large organs. CT-SAM3D significantly outperforms all previous SAM-derived models by a large margin and demonstrates strong zero-shot capability.

References

AlBadawy, E. A.; Saha, A.; and Mazurowski, M. A. 2018. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Medical physics*, 45(3): 1150–1158.

Antonelli, M.; Reinke, A.; Bakas, S.; Farahani, K.; Kopp-Schneider, A.; Landman, B. A.; Litjens, G.; Menze, B.; Ronneberger, O.; Summers, R. M.; et al. 2022. The medical segmentation decathlon. *Nature communications*, 13(1): 4128.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *NeurIPS*, 33: 1877–1901.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *ICCV*, 9650–9660.

Chen, C.; Miao, J.; Wu, D.; Zhong, A.; Yan, Z.; Kim, S.; Hu, J.; Liu, Z.; Sun, L.; Li, X.; Liu, T.; Heng, P.-A.; and Li, Q. 2024. MA-SAM: Modality-agnostic SAM adaptation for 3D medical image segmentation. *Medical Image Analysis*, 98: 103310.

Chen, X.; Zhao, Z.; Zhang, Y.; Duan, M.; Qi, D.; and Zhao, H. 2022. Focalclick: Towards practical interactive image segmentation. In *CVPR*, 1300–1309.

Cheng, J.; Ye, J.; Deng, Z.; Chen, J.; Li, T.; Wang, H.; Su, Y.; Huang, Z.; Chen, J.; Jiang, L.; Sun, H.; He, J.; Zhang, S.; Zhu, M.; and Qiao, Y. 2023. SAM-Med2D. *arXiv preprint arXiv:2308.16184*.

Chilamkurthy, S.; Ghosh, R.; Tanamala, S.; Biviji, M.; Campeau, N. G.; Venugopal, V. K.; Mahajan, V.; Rao, P.; and Warier, P. 2018. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *The Lancet*, 392(10162): 2388–2396.

Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S. S.; Brox, T.; and Ronneberger, O. 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *MICCAI*, 424–432. Springer.

Deng, R.; Cui, C.; Liu, Q.; Yao, T.; Remedios, L. W.; Bao, S.; Landman, B. A.; Wheless, L. E.; Coburn, L. A.; Wilson, K. T.; et al. 2023. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging. *arXiv preprint arXiv:2304.04155*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, Y.; Bai, F.; Huang, T.; and Zhao, B. 2024. SegVol: Universal and Interactive Volumetric Medical Image Segmentation. *NeurIPS*.

Ferré, E. M.; Break, T. J.; Burbelo, P. D.; Allgäuer, M.; et al. 2019. Lymphocyte-driven regional immunopathology in pneumonitis caused by impaired central immune tolerance. *Science translational medicine*, 11(495): eaav5597.

Gong, S.; Zhong, Y.; Ma, W.; Li, J.; Wang, Z.; Zhang, J.; Heng, P.-A.; and Dou, Q. 2024. 3DSAM-adapter: Holistic adaptation of SAM from 2D to 3D for promptable tumor segmentation. *Medical Image Analysis*, 98: 103324.

Guo, H.; Zhang, J.; Yan, K.; Lu, L.; and Xu, M. 2024. Medquery: Steerable parsing of 9-dof medical anatomies with query embedding. *IEEE Journal of Biomedical and Health Informatics*.

He, S.; Bao, R.; Li, J.; Grant, P. E.; and Ou, Y. 2023. Accuracy of segment-anything model (sam) in medical image segmentation tasks. *arXiv preprint arXiv:2304.09324*.

Hesamian, M. H.; Jia, W.; He, X.; and Kennedy, P. 2019. Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of digital imaging*, 32: 582–596.

Huang, Y.; Yang, X.; Liu, L.; Zhou, H.; Chang, A.; Zhou, X.; Chen, R.; Yu, J.; Chen, J.; Chen, C.; et al. 2024. Segment anything model for medical images? *Medical Image Analysis*, 92: 103061.

Isensee, F.; Jaeger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2): 203–211.

Iyer, K. S.; Newell Jr, J. D.; Jin, D.; Fuld, M. K.; Saha, P. K.; Hansdottir, S.; and Hoffman, E. A. 2016. Quantitative dual-energy computed tomography supports a vascular etiology of smoking-induced inflammatory lung disease. *American*

journal of respiratory and critical care medicine, 193(6): 652–661.

Ji, Z.; Shen, Y.; Ma, C.; and Gao, M. 2019. Scribble-based hierarchical weakly supervised learning for brain tumor segmentation. In *MICCAI*, 175–183. Springer.

Jin, D.; Guo, D.; Ge, J.; Ye, X.; and Lu, L. 2022. Towards automated organs at risk and target volumes contouring: Defining precision radiation therapy in the modern era. *Journal of the National Cancer Center*.

Jin, D.; Guo, D.; Ho, T.-Y.; Harrison, A. P.; Xiao, J.; Tseng, C.-K.; and Lu, L. 2021. DeepTarget: Gross tumor and clinical target volume segmentation in esophageal cancer radiotherapy. *Medical Image Analysis*, 68: 101909.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollar, P.; and Girshick, R. 2023. Segment Anything. In *ICCV*, 4015–4026.

Koohbanani, N. A.; Jahanifar, M.; Tajadin, N. Z.; and Rajpoot, N. 2020. NuClick: a deep learning framework for interactive segmentation of microscopic images. *Medical Image Analysis*, 65: 101771.

Landman, B.; Xu, Z.; Igelsias, J.; Styner, M.; Langerak, T.; and Klein, A. 2015. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, 12.

Lee, H. H.; Bao, S.; Huo, Y.; and Landman, B. A. 2022. 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. *arXiv preprint arXiv:2209.15076*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *ICCV*, 2980–2988.

Liu, Q.; Xu, Z.; Bertasius, G.; and Niethammer, M. 2023. SimpleClick: Interactive image segmentation with simple vision transformers. In *ICCV*, 22290–22300.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024. Segment anything in medical images. *Nature Communications*, 15(1): 654.

Ma, J.; Zhang, Y.; Gu, S.; Ge, C.; Ma, S.; Young, A.; Zhu, C.; Meng, K.; Yang, X.; Huang, Z.; et al. 2023. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. *arXiv preprint arXiv:2308.05862*.

Maciej, A.; Haoyu, D.; Hanxue, G.; Jichen, Y.; Nicholas, K.; and Yixin, Z. 2023. Segment anything model for medical image analysis: an experimental study. *arXiv preprint arXiv:2304.10517*, 2.

Maier-Hein, L.; Menze, B.; et al. 2022. Metrics reloaded: Pitfalls and recommendations for image analysis validation. *arXiv preprint arXiv:2206.01653*.

Maleike, D.; Nolden, M.; Meinzer, H.-P.; and Wolf, I. 2009. Interactive segmentation framework of the medical imaging interaction toolkit. *Computer methods and programs in biomedicine*, 96(1): 72–83.

McKinney, S. M.; Sieniek, M.; Godbole, V.; Godwin, J.; Antropova, N.; Ashrafian, H.; Back, T.; Chesus, M.; Corrado, G. S.; Darzi, A.; et al. 2020. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788): 89–94.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 565–571. IEEE.

Mitani, A.; Huang, A.; Venugopalan, S.; Corrado, G. S.; Peng, L.; Webster, D. R.; Hammel, N.; Liu, Y.; and Varadarajan, A. V. 2020. Detection of anaemia from retinal fundus images via deep learning. *Nature Biomedical Engineering*, 4(1): 18–27.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2337–2346.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Rahimi, A.; and Recht, B. 2007. Random features for large-scale kernel machines. *NeurIPS*, 20.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241. Springer.

Roth, H. R.; Lu, L.; Liu, J.; Yao, J.; Seff, A.; Cherry, K.; Kim, L.; and Summers, R. M. 2015. Improving computer-aided detection using convolutional neural networks and random view aggregation. *IEEE transactions on medical imaging*, 35(5): 1170–1181.

Sakinis, T.; Milletari, F.; Roth, H.; Korfiatis, P.; Kostandy, P.; Philbrick, K.; Akkus, Z.; Xu, Z.; Xu, D.; and Erickson, B. J. 2019. Interactive segmentation of medical images through fully convolutional neural networks. *arXiv preprint arXiv:1903.08205*.

Shaker, A.; Maaz, M.; Rasheed, H.; Khan, S.; Yang, M.-H.; and Khan, F. S. 2022. UNETR++: delving into efficient and accurate 3D medical image segmentation. *arXiv preprint arXiv:2212.04497*.

Shin, H. C.; Roth, H. R.; Gao, M.; Lu, L.; Xu, Z.; Nogues, I.; Yao, J.; Mollura, D.; and Summers, R. M. 2016. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging*, 35(5): 1285–1298.

Sofiiuk, K.; Petrov, I. A.; and Konushin, A. 2022. Reviving iterative training with mask guidance for interactive segmentation. In *ICIP*, 3141–3145. IEEE.

Tancik, M.; Srinivasan, P.; Mildenhall, B.; Fridovich-Keil, S.; Raghavan, N.; Singhal, U.; Ramamoorthi, R.; Barron, J.; and Ng, R. 2020. Fourier features let networks learn high frequency functions in low dimensional domains. *NeurIPS*, 33: 7537–7547.

Tang, Y.; Yang, D.; Li, W.; Roth, H. R.; Landman, B.; Xu, D.; Nath, V.; and Hatamizadeh, A. 2022. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *CVPR*, 20730–20740.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

Wachinger, C.; Toews, M.; Langs, G.; Wells, W.; and Golland, P. 2018. Keypoint transfer for fast whole-body segmentation. *IEEE transactions on medical imaging*, 39(2): 273–282.

Wald, T.; Roy, S.; Koehler, G.; Disch, N.; Rokuss, M. R.; Holzschuh, J.; Zimmerer, D.; and Maier-Hein, K. 2023. SAM. MD: Zero-shot medical image segmentation capabilities of the Segment Anything Model. In *Medical Imaging with Deep Learning, short paper track*.

Wang, G.; Li, W.; Zuluaga, M. A.; Pratt, R.; Patel, P. A.; Aertsen, M.; Doel, T.; David, A. L.; Deprest, J.; Ourselin, S.; et al. 2018. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE transactions on medical imaging*, 37(7): 1562–1573.

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.; and Qiao, Y. 2023a. SAM-Med3D. *arXiv preprint arXiv:2310.15161*.

Wang, X.; Zhang, L.; Roth, H.; Xu, D.; and Xu, Z. 2019. Interactive 3D segmentation editing and refinement via gated graph neural networks. In *International Workshop on Graph Learning in Medical Imaging*, 9–17. Springer.

Wang, X.; Zhang, X.; Cao, Y.; Wang, W.; Shen, C.; and Huang, T. 2023b. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*.

Wasserthal, J.; Breit, H.-C.; Meyer, M. T.; Pradella, M.; Hinck, D.; Sauter, A. W.; Heye, T.; Boll, D. T.; Cyriac, J.; Yang, S.; et al. 2023. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5).

Wu, J.; Fu, R.; Fang, H.; Liu, Y.; Wang, Z.; Xu, Y.; Jin, Y.; and Arbel, T. 2023. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*.

Xu, N.; Price, B.; Cohen, S.; Yang, J.; and Huang, T. S. 2016. Deep interactive object selection. In *CVPR*, 373–381.

Ye, X.; Guo, D.; Ge, J.; Yan, S.; Xin, Y.; Song, Y.; Yan, Y.; Huang, B.-s.; Hung, T.-M.; Zhu, Z.; et al. 2022. Comprehensive and clinically accurate head and neck cancer organs-at-risk delineation on a multi-institutional study. *Nature Communications*, 13(1): 6137.

Yue, W.; Zhang, J.; Hu, K.; Xia, Y.; Luo, J.; and Wang, Z. 2023. SurgicalSAM: Efficient Class Promptable Surgical Instrument Segmentation. *arXiv preprint arXiv:2308.08746*.

Zhang, J.; Shi, Y.; Sun, J.; Wang, L.; Zhou, L.; Gao, Y.; and Shen, D. 2021. Interactive medical image segmentation via a point-based interaction. *Artificial Intelligence in Medicine*, 111: 101998.

Zhang, K.; and Liu, D. 2023. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*.

Zhang, Q.; and Yang, Y.-B. 2022. Rest v2: simpler, faster and stronger. *NeurIPS*, 35: 36440–36452.

Zhao, F.; and Xie, X. 2013. An overview of interactive medical image segmentation. *Annals of the BMVA*, 2013(7): 1–22.

Zhou, H.-Y.; Guo, J.; Zhang, Y.; Han, X.; Yu, L.; Wang, L.; and Yu, Y. 2023. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE Transactions on Image Processing*.

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Gao, J.; and Lee, Y. J. 2023. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*.