VISTA3D: A Unified Segmentation Foundation Model For 3D Medical Imaging

Yufan He[1] , Pengfei Guo[1] , Yucheng Tang[1] , Andriy Myronenko[1] , Vishwesh Nath[1],
Ziyue Xu[1], Dong Yang[1] , Can Zhao[1] , Benjamin Simon[3,4] , Mason Belue[2],
Stephanie Harmon[3] , Baris Turkbey[3] , Daguang Xu[1] , Wenqi Li[1]
[1] NVIDIA
[2] University of Arkansas for Medical Sciences
[3]National Institutes of Health [4] University of Oxford

Abstract

*Foundation models for interactive segmentation in 2D natural images and videos have sparked significant interest in building 3D foundation models for medical imaging. However, the domain gaps and clinical use cases for 3D medical imaging require a dedicated model that diverges from existing 2D solutions. Specifically, such foundation models should support a full workflow that can actually reduce human effort. Treating 3D medical images as sequences of 2D slices and reusing interactive 2D foundation models seems straightforward, but 2D annotation is too time-consuming for 3D tasks. Moreover, for large cohort analysis, it's the highly accurate **automatic** segmentation models that reduce the most human effort. However, these models lack support for interactive corrections and lack zero-shot ability for novel structures, which is a key feature of "foundation". While reusing pre-trained 2D backbones in 3D enhances zero-shot potential, their performance on complex 3D structures still lags behind leading 3D models. To address these issues, we present VISTA3D, **V**ersatile **I**maging **S**egmen**T**ation and **A**nnotation model, that targets to solve all these challenges and requirements with one unified foundation model. VISTA3D is built on top of the well-established 3D segmentation pipeline, and it is the first model to achieve state-of-the-art performance in both 3D automatic (supporting 127 classes) and 3D interactive segmentation, even when compared with top 3D expert models on large and diverse benchmarks. Additionally, VISTA3D's 3D interactive design allows efficient human correction, and a novel 3D supervoxel method that distills 2D pre-trained backbones grants VISTA3D top 3D zero-shot performance. We believe the model, recipe, and insights represent a promising step towards a clinically useful 3D foundation model. Code and weights are publicly available at* https://github.com/Project-MONAI/VISTA.

1. Introduction

Three-dimensional medical imaging such as computed tomography (CT) is widely used for creating cross-sectional volumetric images within various body regions. As a major anatomic imaging modality, it reveals detailed morphological information of body structures and abnormalities. In clinical practice, manual segmentation is time-consuming and tedious, thus developing better automatic models has been one of the most active research topics. A typical direction is enhancing network architecture and tailoring training recipes for specific tasks [18, 22, 37, 54]. For each task, curating a specific set of training data and training expert models is often performed, which requires strong engineering expertise. A model that solves a variety of tasks out of the box is thus more desirable.

Unlike natural images where there could be an unlimited number of object classes, the clinically relevant healthy human anatomies revealed by CT or MRI are limited (such as liver, pancreas), thus training an **automated** segmentation model that supports most of standard human anatomies is technically feasible [12, 24, 29, 58, 61]. However, in practice, clinicians may be more interested in rare pathologies or animal data that are usually unsupported by those models due to data scarcity. Lacking zero-shot capability to handle those use cases becomes a significant limitation. Meanwhile, it is important for the model to allow human input for correction for procedures like surgical planning.

Recently, large language models [2, 55, 57] have shown strong generalizability on various tasks and are considered the foundation models. The idea of a "promptable" system has been proposed to achieve a flexible model that can solve different tasks out-of-the-box. For image segmentation, Segment Anything (SAM) [26] has gained great interest and achieved impressive zero-shot performance. In the medical domain, recent work [33] hence adapted SAM to medical imaging modalities via model fine-tuning. These SAM-based methods demonstrate promising results in 2D

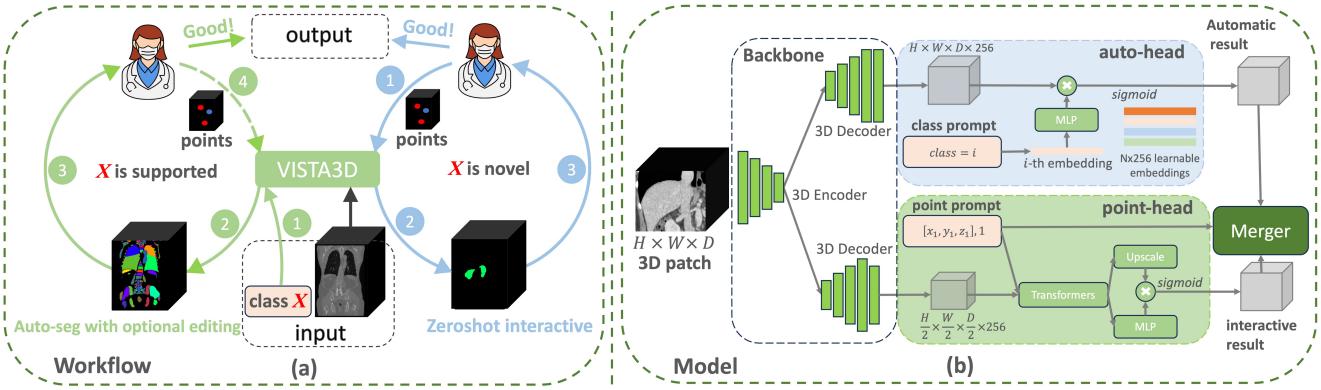

Figure 1. Fig.(a) shows the full human-in-the-loop workflow VISTA3D supports. If the segmentation task X is within 127 supported classes (left green circle), VISTA3D performs accurate automatic segmentation. The doctor can inspect and efficiently edit the result with VISTA3D if needed. If X is a novel class (right blue circle), VISTA3D performs 3D interactive zero-shot segmentation. Fig.(b) shows the VISTA3D architecture. It contains two branches that share the same image encoder. The top auto-branch will activate out-of-the-box automatic segmentation if user provide a class prompt that's within 127 supported classes. The bottom interactive branch will activate interactive segmentation if user provide 3D point click prompts. If both branches are activated, a merger module based on Alg. 1 will use interactive results to edit automatic results.

that leverage interactive user input. For 3D medical images, such prompt (e.g. point) binding to every class, every slice, and every scan, often requires substantial human effort, making it infeasible for large cohort data analysis. The recent Segment Anything in Video (SAM2) [44] triggered even greater interest since a 3D scan is represented by a stack of 2D cross-sectional images (slices), while a video is also a stack of 2D images (frames). However, our experiments show that the SAM2 framework, even well-finetuned on 3D medical datasets, cannot compare with VISTA3D, especially on complicated 3D structures (details in the supplementary). This illustrates the big gap between 2D natural images or videos and cross-sectional medical images. Similarly, SAM3D [5] extracts 3D volume features slice by slice with a 2D SAM encoder and a 3D decoder, but the results are much worse than 3D experts. Simply applying methods from natural images to 3D medical images will fall short.

Recent works exploring in-context learning for medical image segmentation [6, 45, 51, 59] can segment any class guided by example image or text. This seems like an optimal case because it does not require model finetuning or time-consuming human input. However, the performance of such methods is far behind [45] the dataset-specific supervised models (e.g. nnU-Net [22]).

We envision that a foundation model for 3D medical image segmentation should support a full workflow (Fig. 1(a)) that can reduce human effort, which may have the following essential capabilities: 1) Highly accurate automatic segmentation for common organs or structures; 2) Ability to interact with human experts, allowing for effective refinements of existing segmentation results; 3) Zero-shot capabilities, either allow the user to interactively annotate unseen classes or use in-context learning via text or example guidance. The model should operate in 3D since 2D slice-by-slice methods are too time-consuming and may not fully leverage 3D visual contexts; 4) Few shot/transfer learning abilities that allow users to quickly finetune the model to perform accurate automatic segmentation on new classes, given that existing in-context learning or open vocabulary segmentation still fall short compared to expert 3D models in accuracy.

To support this workflow and achieve comparable performances with the best expert models, we should build the model based on well-established 3D pipelines that rely on 3D backbones and sliding window inference. However, this direction is not taking advantage of the existing 2D pretrained weights with a strong zero-shot ability(e.g. SAM). Reusing SAM weights and adding light-weight 3D adaptor modules [14, 60] seems viable but the automatic performance on diverse classes (comparing to TotalSegmentator [12]) are limited due to freezing the majority of weights. So the challenge is, how to build a model that possesses the advantage of well-established 3D pipelines, while also utilizing the insights and checkpoints from 2D natural images to solve 3D problems. Based on this goal, we introduce VISTA3D, and our contributions are:

1. The first unified foundation model that supports a full annotation workflow, and achieves state-of-the-art 3D promptable automatic segmentation and interactive editing, benchmarked over 14 challenging datasets with 127 classes and compared with well-established baselines.
2. A novel supervoxel methods are developed to distill 2D foundation models for 3D medical imaging, which boosted VISTA3D's zero-shot performance by 50% and

achieved state-of-the-art 3D zero-shot performance with much less annotation efforts.

3. We curated a large CT dataset with 11454 scans, paired it with partial manual labels, pseudo labels, and supervoxels, and proposed a novel four-stage training recipe to tackle the challenges to achieve state-of-the-art performances and editing experiences.

2. Related Work

Dataset-specific supervised training. Many existing 3D medical imaging segmentation methods [18, 22, 37, 54] are proposed to train dedicated models for a specific dataset. nnU-Net [22] is a well-established framework that can automatically adapt to different datasets. This adaptation occurs seamlessly, mitigating the need for manual intervention or specialized expertise, which expedites the medical imaging segmentation models. Auto3DSeg[1] presents a holistic approach to tackling the challenge of large-scale 3D medical image segmentation. It also provides automatic task adaptation. These two frameworks have proven their effectiveness by winning numerous highly competitive 3D segmentation challenges [10, 13, 22, 38–43]. Although auto-configuration solutions can speed up the curation of task-specific expert models and achieve high performance, they lack inherent zero-shot capabilities and require human effort and resource in data preparation and training.

Foundational segmentation model. Foundation segmentation models aim to develop a single unified deep learning model capable of segmenting multiple anatomical structures/organs from whole-body CT scans, rather than training separate models for each organ. Totalsegmentator [12] is proposed for fully automatic segmentation of over 117 anatomical structures in CT images covering various organs, bones, muscles, and vessels. It represents a significant contribution to the biomedical imaging community, enabling researchers and clinicians to leverage accurate and comprehensive segmentation without requiring time-consuming manual efforts. The Universal Model [29] leverages text embeddings from the CLIP model to encode the anatomical relationships between organs and tumors and support the automatic segmentation of 31 classes. SAT [61] supports automatic segmentation on 497 classes based on text prompts. Although they tried to incorporate text embeddings into automatic segmentation, the text prompts only work with a fixed vocabulary and do not support zero-shot or open-vocabulary segmentation. Continual Segment [24] is a unified model capable of segmenting 143 body organs by a frozen encoder coupled with incrementally added decoders to avoid catastrophically forgetting previously learned structures. While those foundational 3D segmentation models represent significant advancements in multi-organ segmentation, the inability of zero-shot and interactive segmentation impedes their real-world applicability. The in-context learning or open vocabulary segmentation [6, 45, 51, 59] are desirable, which can achieve automatic segmentation on the unseen class by prompts or support examples. However, at the current stage, their performances fall short compared to the expert models, especially for 3D images [45].

Interactive medical image segmentation. The Segment Anything Model (SAM) [26] and its video variant (SAM2) [44] have inspired and enabled various medical imaging applications through the adaptation and finetuning of medical data [14, 21, 33, 60]. MedSAM [33] finetunes SAM on large 2D medical datasets using the bounding box but lacks the ability for detailed editing and handling 3D inputs. The SAM adapters [14, 60] add 3D adaptor modules to the SAM backbone for efficient 3D finetuning, however, the 3D performance was validated on limited classes and benchmarks. The work closely related to ours is SegVol [11], a 3D foundation model designed for 3D semantic and interactive segmentation. However, SegVol [11]'s performance relies heavily on the 3D bounding box added with text prompts, and there is still a big performance gap with its text prompt-based automatic segmentation. Although those work approved their effectiveness in segmenting 10 to 20 classes of 3D structures on benchmarks like BTCV [28] or AMOS [23], the problem is, that those structures can be easily solved by automatic foundation models like TotalSegmentator [12]. It is important to rethink the position of interactive models and how they can really reduce human efforts for 3D medical segmentation.

3. Method

3.1. Overview

We separate the segmentation tasks into **supported classes** and **zero-shot classes**. The supported classes are the classes that have enough training data with annotations with which we can train VISTA3D to perform automatic segmentation (here we support 127 classes). We curated a global class index list and mapped the groundtruth indices from those partially annotated datasets to this list. We trained the automatic head to accept the index as the prompt and output a binary segmentation. For zero-shot classes, the segmentation is mainly generated by the VISTA3D interactive branch, which accepts user click coordinates in 3D. The interactive segmentation also works for supported classes. The overall workflow is shown in Fig. 1(a). To train such a model, we curated a large dataset containing 11454 3D CT scans, generated pseudo labels from TotalSegmentator model [12] and supervoxels using SAM pre-trained weights [26](see detail in Sec 3.3). A stage-by-stage training recipe is used to train interactive and automatic workflows systematically.

[1] https://monai.io/apps/auto3dseg

3.2. Model architecture

SAM's [26] image encoder is a vision transformer (ViT) [27] with 16×16 patch embedding. For 3D images, ViT becomes extremely memory-demanding, since the token length (number of patches) is much longer compared to 2D images. On the other hand, 16×16 patch embedding inevitably loses spatial details. Computationally feasible adaptations of transformers to 3D images have been proposed [17, 19], but the state-of-the-art results, as shown in the recent MICCAI 3D segmentation challenges, are still predominantly based on the convolutional architectures. Specifically, the SegResNet model, a U-net type architecture, has won BraTS 2023 [37], KiTS 2023 [40] and Seg.A 2023 [42] MICCAI 3D segmentation challenges. In VISTA3D, we use SegResNet [37] from MONAI [7], as a backbone CNN, and followed the best practices in medical image segmentation of patch-based training (we used 128-voxel cubic patch) and sliding window inference.

Automatic branch As shown in Fig. 1(b), there are two branches: the automatic branch (top) and the interactive branch (bottom). The SegResNet encoder is shared between two branches for learning image embedding. Each branch has its own decoder with a skip connection with the shared encoder. The auto-head contains an MLP layer M and a learnable $N \times C$ class embedding E_c (we use $C = 256$), where N represents N supported classes. The output feature F from the decoder is of size $C \times H \times W \times D$. If the user wants to segment class i (this single number i is the input prompt for the auto-head), the corresponding class embedding $E_c[i]$ is used to map the feature into segmentation logits, $sigmoid(M(E_c[i]) \times F)$. Compared to models that output all classes and apply $softmax$ at the end, our promptable scheme reduces memory usage dramatically if number of classes are huge. Meanwhile, it avoids the partial label problem in training on diverse partially labeled datasets. We added this additional MLP layer M due to empirically better performance.

Works [11, 29, 61] have tried to use text embedding like CLIP but none of them are able to achieve zero-shot or open vocabulary segmentation with text, which is the main benefit of text prompts. Moreover, we empirically found CLIP-embedding gives slightly worse results than randomly initialized class embeddings (used by VISTA3D).

Interactive branch For the interactive branch, the click points' 3D coordinates and their labels (positive or negative) are accepted as prompts for the point head. The point head is based on the SAM's [26] point prompt encoder, where the feature map performs cross-attention with point embedding. We made several changes to satisfy the needs for 3D medical images: 1) To keep the high-resolution details, the point head input feature is first upsampled back to the original image resolution with the long skip connection and then 2x downsampled to reduce the memory footprint. All the related operations including point embedding are changed to 3D. The downsampled feature and the point embedding will go through cross-attention transformers and generate the final output. 2) To increase the click response speed for a better user experience, only a local patch centered at the click point will be segmented and used to refine automatic results. 3) For some classes that have ambiguity or overlap, e.g. pancreas/pancreas tumor and colon/colon tumor, a single point click cannot solve the ambiguity. Since the model knows the class x to segment beforehand, we can add a special embedding to the click point automatically if x lies in specific classes like colon/pancreas tumor. This embedding can be used to distinguish ambiguous classes. Note that this special embedding is the same for all classes with ambiguity. If we use class embedding in E_c to solve the ambiguity, the point head will learn a shortcut to ignore point clicks. 4) Another challenge is that the interactive branch needs to handle both supported classes and unseen classes (zero-shot), while there might be a conflict between these two tasks. Segmenting supported class with high accuracy will require the model to remember or overfit specific features about the class like the shape and position. However, organ-specific tuning could hurt zero-shot generalizability. We mitigate this problem by adding a zero-shot embedding to the point head cross-attention if the class x is a novel class. Examples can be found in the supplementary.

Interactive refinement over automatic results As can be seen in Fig. 1(b), the two branch outputs are independent from each other. The use case by combining the results is to interactively correct the automatic segmentation results. As illustrated in FocalClick [8], the interactive refinement over existing masks could destroy the correct part. We observe this behavior when simply combining the interactive results and the automatic results. We used the local refinement idea from FocalClick and proposed the following merging algorithm Alg. 1. The core idea is to add or remove only the connected component regions that contain the point clicks to avoid unexpected modification.

3.3. Data

We curated a collection of 11454 CT volumetric images obtained from in-house and publicly available data sources [3, 4, 12, 16, 20, 23, 25, 31, 32, 46–50, 52, 53, 56] with a wide range of acquisition protocols and subject conditions. Among these, five of them are without labels, and the rest have various voxel-wise annotation regions of interest, including anatomical structures and lesions. We denote the ground truth from those datasets as **manual labels** or **partial labels**. Each data source is randomly split into 64% training, 16% validation, and 20% test sets. We generated **pseudo-labels** of 117 classes using TotalSegmentator [12] and **supervoxels** using SAM for every scan. The unreliable pseudo-labels are removed by post-processing.

Algorithm 1 Interactive refinement on the automatic results

Require: I positive and J negative clicks P_p^i and P_n^j, automatic and interactive output M_a and M_p

Ensure: $size(M_a) = size(M_p)$

Denote "get 3D connected components" as $CC(\cdot)$.

$\{M_{add}^n\}_N \leftarrow CC((M_p - M_a) > 0)$ ▷ N added connected components

$\{M_{rm}^k\}_K \leftarrow CC((M_a - M_p) > 0)$ ▷ K removed connected components

if $\exists_{i=1,\cdots,I} P_p^i \in M^a$ **then** $M_{add}^n = M_{add}^n \cup CC(M^p)$

end if ▷ If positive points in M_a, add M_p into addition candidates.

$M_{final\_rm}, M_{final\_add} \leftarrow \{\}, \{\}$

for $n = 1$ to N **do**

if $\exists_{i=1,\cdots,I} P_p^i \in M_{add}^n$ **then** $M_{final\_add} = M_{final\_add} \cup M_{add}^n$

end if

end for

for $k = 1$ to K **do**

if $\exists_{j=1,\cdots,J} P_n^j \in M_{rm}^k$ **then** $M_{final\_rm} = M_{final\_rm} \cup M_r^k m$

end if

end for

return $M_a + M_{final\_add} - M_{final\_rm}$

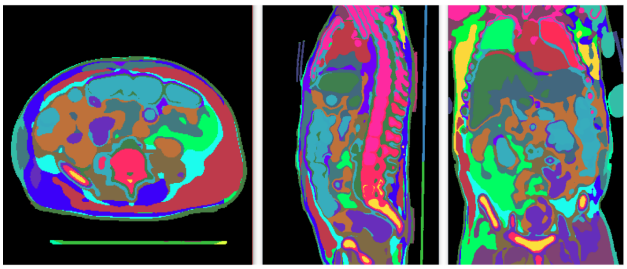

Figure 2. Generated supervoxel from Alg. 2, showing examples in axial, sagittal, and coronal views. Different colours represent different supervoxels.

Supervoxel generation The vast majority of SAM's zero-shot capabilities come from this large-scale supervised training on its 11 million diverse and fully annotated images [26]. Those annotations helped SAM learn how humans perceive an object, and become the image segmentation foundation model. However, the manual labels or pseudo-labels in 3D CT can only cover around one or two hundred classes. We empirically found that this level of class diversity is not enough for the model to achieve SAM-like zero-shot ability in 3D. To solve this problem, most works decided to finetune SAM pretrained ViT checkpoint on 2D medical data to inherit this zero-shot ability, which inevitably limited the adaptability to 3D images. Here we propose a novel method to distill the image understanding ability from SAM by generating 3D supervoxels from 2D SAM feature maps. The algorithm is shown in Alg. 2. We perform a 3D supervoxel algorithm on the upsampled SAM feature embedding, which is generated slice-by-slice in three views. An example of generated supervoxel results is shown in Fig. 2. We generate supervoxels for all 11454 CT scans and use them to train our interactive branch, and this gives VISTA3D zero-shot capabilities. SegVol [11] used a similar idea but the supervoxel generation is based on graph-cut, which is still on low-level image features. Instead, VISTA3D achieved better zero-shot performance through distilling knowledge from SAM.

Algorithm 2 3D supervoxel generation from SAM

Require: SAM pretrained ViT-H model Φ, image encoder Φ_E, output scaling layer in the mask decoder Φ_s. ▷ All SAM components related to prompts are removed

Ensure: Input 3D CT image V

$V \leftarrow \{x_1, x_2, \cdots, x_A\}$ ▷ V as a stack of axial slices

$V \leftarrow \{y_1, y_2, \cdots, y_C\}$ ▷ V as a stack of coronal slices

$V \leftarrow \{z_1, z_2, \cdots, z_S\}$ ▷ V as a stack of sagittal slices

$F_A, F_C, F_S \leftarrow \{\}, \{\}, \{\}$

for $i = 1$ to $A; j = 1$ to $C; k = 1$ to S **do**

$F_A = F_A \cup \Phi_S(\Phi_E(x_i))$

$F_C = F_C \cup \Phi_S(\Phi_E(y_j))$

$F_S = F_S \cup \Phi_S(\Phi_E(z_k))$ ▷ Generate upsampled SAM feature for each slice at each axis.

end for

$F_{3D} \leftarrow F_A + F_C + F_S$ ▷ F_A, F_C, F_S are 3D tensors with the same size

return $SLIC(F_{3D}, n_{segments} = 100, sigma = 3)$ ▷ We use SLIC [1] algorithm from skimage

3.4. Recipe

The training has four stages to solve the class imbalance issues and complications between the automatic and interactive branches.

Stage1-Interactive branch training: This is the first stage of VISTA3D training, and the goal is to train a strong image encoder that can extract good and generalizable features from 3D CT images, and enable the interactive branch to have good response to point clicks. In each iteration, the inputs contain randomly cropped 128 cubic image patches, corresponding manual labels, pseudo labels, and supervoxels. We use a point sampler (details in supplementary) to randomly sample points and its corresponding binary segmentation mask for training. The mask is generated by combining manual labels or pseudo labels with supervoxels or supervoxels alone. The goal is to diversify the ground truth and make the model responsive to all kinds of objects and boundaries. We also followed the SAM's iterative train-

ing scheme and sampled new points from the false positive or negative regions from previous predictions to improve editing ability. We used an iteration of 5.

Stage2-Interactive branch finetuning: The data imbalance issue is severe in our curated dataset since some rare classes such as tumors are only presented in a limited number of images. Stage 1 has a large number of training iterations and if we oversample under-represented classes during stage 1, those classes will soon overfit, moreover, overfit at different iterations for different classes. We disabled oversampling in stage 1, and under-represented classes were rarely sampled (still needs to be sampled). In stage 2, we perform a quick finetuning with specific dataset oversampling to improve low-performing classes. Meanwhile, we removed the supervoxel and unlabeled dataset in finetuning.

Stage3-Automatic branch training: The image encoder has been trained in the previous stages with all the 3D medical annotations and SAM-generated supervoxels. The training is based on binary segmentation without any class-specific information thus we can expect the encoder to generate more generalizable features. We freeze the image encoder to avoid changes to the interactive branch and train the auto branch decoder and head. The training is a common supervised training, but we randomly sample existing class indexes from the manual labels and pseudo-labels and use the corresponding binary masks as ground truth.

Stage4-Automatic branch finetuning: Similar to stage 2, we need to improve the performance of under-represented classes. We used MAISI [15] to generate synthetic data containing anomalies such as tumors and lesions to enlarge the under-represented class sample size. We sample uniformly across dataset and finetuned model for a few epochs.

4. Supported classes results

We first test the performance of the supported classes. For supported classes, the out-of-the-box performance of a foundational model should have state-of-the-art or comparable performances to the data-specific expert models. Meanwhile, we claim VISTA3D interactive branch can correct error regions in automatic results. We show the VISTA3D's out-of-the-box automatic segmentation results (VISTA3D auto), interactive results with a single positive click sampled from the foreground center (VISTA3D point), and the corrected automatic results with a single click point (VISTA3D auto + point) with Alg. 1 in Table. 1. The click point is randomly sampled from the false positive (negative point) or false negative region (positive point), based on which has a larger area size. Note that VISTA3D is a patch-based method using sliding window inference, thus a click point will only affect the 128 cubic patch that includes the point. The evaluation with a single point means 1 click for each sliding window patch. As for baselines, we used the Auto3Dseg framework and the nnUNet framework to train expert models for each dataset (same train/val/test split as VISTA3D) until full convergence. TotalSegmentator is applied out-of-the-box as a foundation model. These three baselines are the well-established "go-to" options for automatic segmentation, and VISTA3D achieved comparable "auto" performances and much better performance if minimum human input is available. Meanwhile, our model is much faster than Totalsegmentator (it has 5 model ensemble) as shown in Table. 2. In Fig. 3, we show an example of using click points to correct automatic results. In Fig. 4, we present a case of automatic segmentation over a **monkey** scan, showing the generalizability of the VISTA3D model. More examples are in the supplementary.

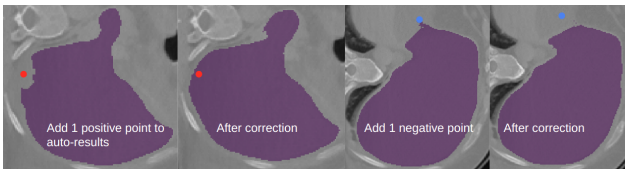

Figure 3. Correcting automatic segmentation with points. The left figure shows the automatic liver segmentation with a false negative area. After a positive point, the false negative region is corrected. The third figure shows another slice with a false positive and a negative point removed from the region shown in the last figure.

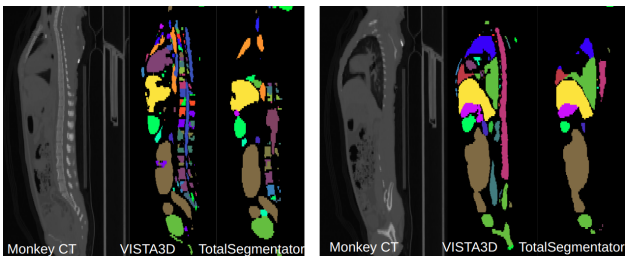

Figure 4. An example of monkey CT scan (2 sagittal slices). We can see that VISTA3D achieved more robust segmentation.

5. Zero-shot results

In this section, we test the zero-shot ability of VISTA3D. We compare with MedSAM [33] and SegVol [11] as they showed the best interactive performances in 2D and 3D separately. For the MedSAM baseline, we adopt the 3D inference pipeline via a series of 2D slices as described in [33]. For segmentation targets that are larger than 10 voxels, tight bounding boxes for each slice were generated to simulate user-provided prompts. Each bounding box is considered the same annotation effort as two-point prompts in our evaluations. For the SegVol baseline, the default settings [11] are evaluated using a positive point with three pairs of positive and negative points (7 points in total), as well as the zoom-out-zoom-in inference strategy. For VISTA3D, We

Table 1. Average dice score of the test split in each dataset. TotalSegV2 results are biased towards nnUNet and TotalSegmentator (the ground truth is generated by the pretrained TotalSegmentator model, which uses nnUNet architecture, and the training data may include our test split. Noted with *). The Bone Lesion is a private dataset with 237 CT scans. Detailed results of all classes are in the supplementary.

	Auto3dSeg	nnUNet	TotalSegmentator	VISTA3D auto	VISTA3D point	VISTA3D auto+point
MSD03 Hepatic tumor [3]	0.616	0.617	-	0.588	**0.701**	0.687
MSD06 Lung tumor [3]	0.562	0.554	-	0.613	0.682	**0.719**
MSD07 Pancreatic tumor [3]	0.485	0.488	-	0.324	0.603	**0.638**
MSD08 Hepatic tumor [3]	0.683	0.659	-	0.682	0.733	**0.757**
MSD09 Spleen [3]	0.965	**0.967**	0.966	0.952	0.938	0.954
MSD10 Colon tumor [3]	0.475	0.473	-	0.439	0.609	**0.633**
Airway [53]	0.896	**0.899**	-	0.852	0.819	0.867
Bone Lesion	0.343	0.396	-	0.491	0.536	**0.585**
BTCV-Abdomen [47]	0.807	0.825	0.846	0.849	0.815	**0.859**
BTCV-Cervix [48]	0.598	0.640	0.611	0.672	0.736	**0.775**
VerSe [50]	0.786	0.828	0.832	0.825	0.896	**0.906**
AbdomenCT-1K [31]	0.934	0.939	0.921	0.935	0.903	**0.940**
AMOS22 [23]	0.854	0.854	0.824	0.841	0.785	**0.856**
TotalSegV2 [12]	0.882	*0.906	***0.942**	0.893	0.884	0.918
Average	0.706	0.718	-	0.711	0.760	**0.792**

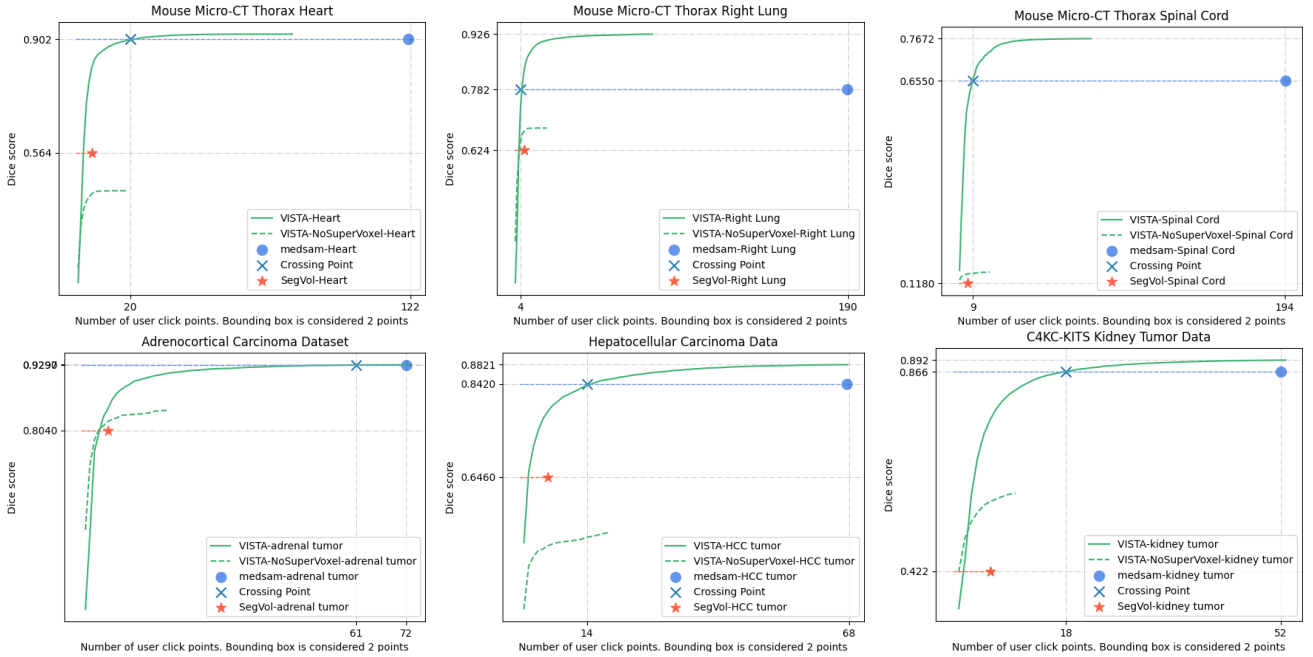

Figure 5. Zero-shot dice scores. The X-axis is the number of click points. The Y-axis is the average dice score over the whole dataset.

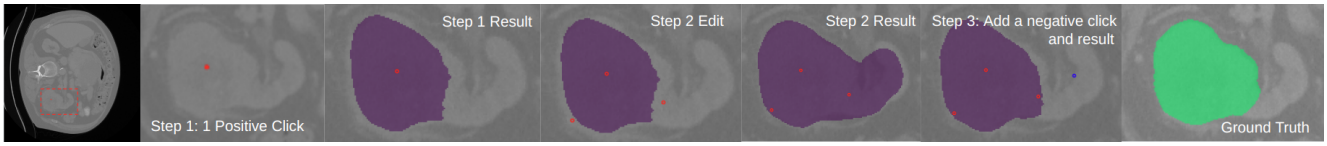

Figure 6. The fine-grained zero-shot interactive segmentation on kidney tumor. The first figure shows the region of the tumor. Step 1 click a positive point (red) on the tumor and get the results. Step 2 click more points to refine the details. The result has over-segmentation and add a negative point (blue) on step 3 to get the final results.

Table 2. Inference speed on a 16GB V100 GPU with varying image sizes (size after resampled to 1.5x1.5x1.5mm). Default 118 class automatic segmentation using VISTA3D and default 117 class segmentation using TotalSegmentator [12]. No special model optimization (e.g. tensorRT).

Size	333x333x603	512x512x512	512x512x768	1024x1024x512
VISTA3D	1m07s	2m09s	3m25s	9m20s
TotalSeg.	4m34s	12m01s	18m56s	40m13s

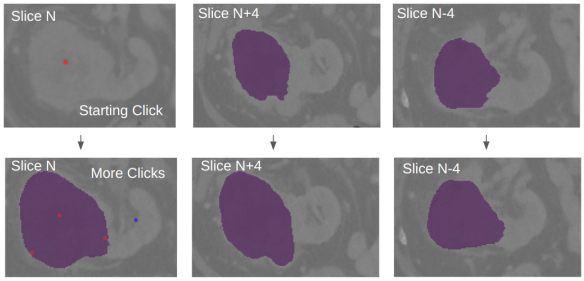

Figure 7. The adjacent slices (slice N+4 and slice N-4) responses to the clicks on slice N (the same slice and click on Fig. 6). The results show how the clicks affect 3D space.

mimic user annotations to perform iterative point clicks. The first point is sampled at the foreground center, then the next point will be randomly sampled from the largest connected false positive or false negative region, which has a larger area size. We evaluate the performance of 4 external datasets with novel classes that our automatic segmentation does not support. 1) the murine dataset [34] includes 140 micro-CT scans (0.2mm resolution) with 4 annotated mouse organs: heart, left lung, right lung, and spinal cord. The left lung showed similar results to the right lung, and the example mouse CT scan and left lung results are shown in the appendix. 2) the C4KC-KITS (kidney tumor, 210 scans) dataset [20], the Adrenocortical Carcinoma (53 scans) dataset [9, 35], the Hepatocellular Carcinoma (105 scans) dataset [9, 36]. The results are shown in Fig. 5. The results show the superior performance of VISTA3D in both accuracy and reduced annotation efforts. VISTA3D trained without supervoxel (VISTA-NoSupervoxel) is also shown in the figure, and the results showed the importance of supervoxel for the zero-shot ability. In Fig. 6, we show the iterative point clicks of a kidney tumor from C4KC-KITS dataset [20]. From step 1 to step 3 we can see how the segmentation responds to the clicks. Fig. 7 shows the responses on other slices without clicks. This shows how the clicks respond in 3D space.

6. Finetuning results

In many cases, interactive annotation is used to curate enough data to train an accurate automatic model. We show the potential of transfer learning using VISTA3D pretrained checkpoint. We perform finetuning on the automatic branch under the setting of one-shot, five-shot, and until the full training data split. The dataset we use includes the Whole abdominal Organs Dataset (WORD) [30] and the micro-CT mouse dataset [34]. We compare with training from scratch methods (nnU-Net and Auto3DSeg, default setting), and finetuning Totalsegmentator pretrained checkpoint with default nnU-Net pipeline. We use the train/val/test data split from WORD [30] and our own split for the mouse dataset. The results of the held-out test set are shown in Table 3. Compared with the baselines, VISTA3D showed much better performance under few shot setting. Meanwhile, the WORD results of using the full training split (100 cases) are directly comparable with all the baselines in WORD [30] paper and VISTA3D has the highest dice score (0.875) over all 10 baselines. The results support our claim where users can annotate a few examples and finetune VISTA3D to build a data flywheel.

Table 3. Finetuning performances of average test dice scores, with respect to the number of training cases.

Dataset	# of cases	Auto3DSeg	nnU-Net	TotalSegmentator	VISTA3D
Micro-CT Mouse	1	0.820	0.759	0.791	**0.926**
	5	0.923	0.922	0.924	**0.935**
	10	0.934	0.930	0.936	**0.938**
	20	**0.947**	0.942	0.944	0.944
	40	**0.949**	**0.949**	**0.949**	0.948
	89	0.949	0.949	**0.951**	**0.951**
WORD	1	0.214	0.185	0.779	**0.795**
	5	0.611	0.562	0.823	**0.839**
	10	0.744	0.697	0.837	**0.855**
	20	0.806	0.793	0.855	**0.862**
	40	0.862	0.831	0.857	**0.869**
	100	0.873	0.874	**0.875**	**0.875**

7. Conclusion

In this paper, we introduced VISTA3D, the first unified 3D CT foundation segmentation model. All the components of VISTA3D are designed to fulfill our proposed human-in-the-loop workflow, such that VISTA3D can be used out-of-the-box to save human effort. It achieved highly accurate segmentation comparable with specialized expert models for each dataset, state-of-the-art interactive segmentation for both zero-shot and results editing, and strong transfer learning ability. The large-scale training data with diverse types of labels, carefully designed model architecture, and training recipes were vital for building this highly capable model. We also utilize the best practices in 3D medical image analysis (e.g. sliding-window, patches, 3D convolutions, data synthesis) to improve the results. For future work, we are working on 1) enlarging the supported class number and modalities, including adding supports for MRI and PET imaging, 2) improving the zero-shot experiences by developing smarter methods to better utilize the datasets and model checkpoints from natural images, and 3) validating and integrating the workflow with clinical partners.

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012. 5

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[3] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022. 4, 7

[4] SG Armato III, G McLennan, L Bidaut, MF McNitt-Gray, CR Meyer, AP Reeves, B Zhao, DR Aberle, CI Henschke, EA Hoffman, et al. Data from LIDC-IDRI [data set]. The Cancer Imaging Archive, 2015. 4

[5] Nhat-Tan Bui, Dinh-Hieu Hoang, Minh-Triet Tran, Gianfranco Doretto, Donald Adjeroh, Brijesh Patel, Arabinda Choudhary, and Ngan Le. Sam3d: Segment anything model in volumetric medical images. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2024. 2

[6] Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Universeg: Universal medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21438–21451, 2023. 2, 3

[7] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022. 4

[8] Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. Focalclick: Towards practical interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1300–1309, 2022. 4

[9] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (TCIA): maintaining and operating a public information repository. *Journal of digital imaging*, 26:1045–1057, 2013. 8

[10] Reuben Dorent, Roya Khajavi, Tagwa Idris, Erik Ziegler, Bhanusupriya Somarouthu, Heather Jacene, Ann LaCasce, Jonathan Deissler, Jan Ehrhardt, Sofija Engelson, et al. Lnq 2023 challenge: Benchmark of weakly-supervised techniques for mediastinal lymph node quantification. *arXiv preprint arXiv:2408.10069*, 2024. 3

[11] Yuxin Du, Fan Bai, Tiejun Huang, and Bo Zhao. Segvol: Universal and interactive volumetric medical image segmentation. *arXiv preprint arXiv:2311.13385*, 2023. 3, 4, 5, 6

[12] Wasserthal et al. TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023. 1, 2, 3, 4, 7, 8

[13] Sergios Gatidis, Marcel Früh, Matthias Fabritius, Sijing Gu, Konstantin Nikolaou, Christian La Fougère, Jin Ye, Junjun He, Yige Peng, Lei Bi, et al. The autopet challenge: towards fully automated lesion segmentation in oncologic pet/ct imaging. 2023. 3

[14] Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation. *arXiv preprint arXiv:2306.13465*, 2023. 2, 3

[15] Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, et al. Maisi: Medical ai for synthetic imaging. *arXiv preprint arXiv:2409.11169*, 2024. 6

[16] Stephanie A Harmon, Thomas H Sanford, Sheng Xu, Evrim B Turkbey, Holger Roth, Ziyue Xu, Dong Yang, Andriy Myronenko, Victoria Anderson, Amel Amalou, et al. Artificial intelligence for the detection of covid-19 pneumonia on chest ct using multinational datasets. *Nature communications*, 11(1):4080, 2020. 4

[17] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3D medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022. 4

[18] Yufan He, Dong Yang, Holger Roth, Can Zhao, and Daguang Xu. Dints: Differentiable neural network topology search for 3d medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5841–5850, 2021. 1, 3

[19] Yufan He, Vishwesh Nath, Dong Yang, Yucheng Tang, Andriy Myronenko, and Daguang Xu. SwinUNETR-V2: Stronger swin transformers with stagewise convolutions for 3D medical image segmentation. In *MICCAI*, 2023. 4

[20] N Heller, N Sathianathen, A Kalapara, E Walczak, K Moore, H Kaluzniak, J Rosenberg, P Blake, Z Rengel, M Oestreich, et al. Data from C4KC-KITS [data set]. *The Cancer Imaging Archive*, 10, 2019. 4, 8

[21] Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, et al. Segment anything model for medical images? *Medical Image Analysis*, 92:103061, 2024. 3

[22] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. 1, 2, 3

[23] Yuanfeng Ji, Haotian Bai, Jie Yang, Chongjian Ge, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, et al. AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *arXiv preprint arXiv:2206.08023*, 2022. 3, 4, 7

[24] Zhanghexuan Ji, Dazhou Guo, Puyang Wang, Ke Yan, Le Lu, Minfeng Xu, Qifeng Wang, Jia Ge, Mingchen Gao, Xi-

anghua Ye, et al. Continual segment: Towards a single, unified and non-forgetting continual segmentation model of 143 whole-body organs in ct scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21140–21151, 2023. 1, 3

[25] C Daniel Johnson, Mei-Hsiu Chen, Alicia Y Toledano, Jay P Heiken, Abraham Dachman, Mark D Kuo, Christine O Menias, Betina Siewert, Jugesh I Cheema, Richard G Obregon, et al. Accuracy of CT colonography for detection of large adenomas and cancers. *New England Journal of Medicine*, 359(12):1207–1217, 2008. 4

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1, 3, 4, 5

[27] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 4

[28] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, page 12, 2015. 3

[29] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21152–21164, 2023. 1, 3, 4

[30] Xiangde Luo, Wenjun Liao, Jianghong Xiao, Jieneng Chen, Tao Song, Xiaofan Zhang, Kang Li, Dimitris N. Metaxas, Guotai Wang, and Shaoting Zhang. Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *Medical Image Analysis*, 82:102642, 2022. 8

[31] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6695–6714, 2021. 4, 7

[32] Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Ma, Adamo Young, Cheng Zhu, Kangkang Meng, Xin Yang, Ziyan Huang, et al. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. *arXiv preprint arXiv:2308.05862*, 2023. 4

[33] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:1–9, 2024. 1, 3, 6

[34] Justin Malimban, Danny Lathouwers, Haibin Qian, Frank Verhaegen, Julia Wiedemann, Sytze Brandenburg, and Marius Staring. Deep learning-based segmentation of the thorax in mouse micro-ct scans. *Scientific reports*, 12(1):1822, 2022. 8

[35] AW Moawad, AA Ahmed, et al. Voxel-level segmentation of pathologically-proven adrenocortical carcinoma with Ki-67 expression (Adrenal-ACC-Ki67-Seg)[data set]. *The Cancer Imaging Archive*, 2023. 8

[36] Ali Morshid, Khaled M Elsayes, Ahmed M Khalaf, Mohab M Elmohr, Justin Yu, Ahmed O Kaseb, Manal Hassan, Armeen Mahvash, Zhihui Wang, John D Hazle, et al. A machine learning model to predict hepatocellular carcinoma response to transcatheter arterial chemoembolization. *Radiology: Artificial Intelligence*, 1(5):e180021, 2019. 8

[37] Andriy Myronenko. 3D MRI brain tumor segmentation using autoencoder regularization. In *MICCAI Brainles. Brain Tumor Segmentation (BraTS) Challenge.*, pages 311–320. Springer, 2018. 1, 3, 4

[38] Andriy Myronenko, Dong Yang, Yufan He, and Daguang Xu. Aorta segmentation from 3d ct in miccai seg. a. 2023 challenge. In *MICCAI Challenge on Segmentation of the Aorta*, pages 13–18. Springer, 2023. 3

[39] Andriy Myronenko, Dong Yang, Yufan He, and Daguang Xu. Automated 3d segmentation of kidneys and tumors in miccai kits 2023 challenge. In *International Challenge on Kidney and Kidney Tumor Segmentation*, pages 1–7. Springer, 2023.

[40] Andriy Myronenko, Dong Yang, Yufan He, and Daguang Xu. Automated 3D segmentation of kidneys and tumors in MICCAI KiTS 2023 challenge. In *MICCAI 20023. International Challenge on Kidney and Kidney Tumor Segmentation*, pages 1–7. Springer, 2023. 4

[41] Andriy Myronenko, Dong Yang, Yufan He, and Daguang Xu. Auto3DSeg for brain tumor segmentation from 3D MRI in BraTS 2023 challenge. In *MICCAI. Brain Tumor Segmentation (BraTS) Challenge.*, 2023.

[42] Andriy Myronenko, Dong Yang, Yufan He, and Daguang Xu. Aorta segmentation from 3D CT in MICCAI SEG.A. 2023 challenge. In *MICCAI 20023. 3D Segmentation of the Aorta Challenge*, 2023. 4

[43] Kelly Payette, Céline Steger, Roxane Licandro, Priscille de Dumast, Hongwei Bran Li, Matthew Barkovich, Liu Li, Maik Dannecker, Chen Chen, Cheng Ouyang, et al. Multicenter fetal brain tissue annotation (feta) challenge 2022 results. *arXiv preprint arXiv:2402.09463*, 2024. 3

[44] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 3

[45] Sucheng Ren, Xiaoke Huang, Xianhang Li, Junfei Xiao, Jieru Mei, Zeyu Wang, Alan Yuille, and Yuyin Zhou. Medical vision generalist: Unifying medical imaging tasks in context. *arXiv preprint arXiv:2406.05565*, 2024. 2, 3

[46] Blaine Rister, Darvin Yi, Kaushik Shivakumar, Tomomi Nobashi, and Daniel L Rubin. CT-ORG, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7(1):381, 2020. 4

[47] H Roth, A Farag, EB Turkbey, L Lu, J Liu, and RM Summers. Data from pancreas-CT (version 2)[data set]. The Cancer Imaging Archive (2016), 2016. 7

[48] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *MICCAI*, pages 556–564. Springer, 2015. 7

[49] Joel Saltz, Mary Saltz, Prateek Prasanna, Richard Moffitt, Janos Hajagos, Erich Bremer, Joseph Balsamo, and Tahsin Kurc. Stony Brook university COVID-19 positive cases. *the cancer imaging archive*, 4, 2021.

[50] Anjany Sekuboyina, Malek E Husseini, Amirhossein Bayat, Maximilian Löffler, Hans Liebl, Hongwei Li, Giles Tetteh, Jan Kukačka, Christian Payer, Darko Štern, et al. Verse: A vertebrae labelling and segmentation benchmark for multi-detector CT images. *Medical image analysis*, 73:102166, 2021. 4, 7

[51] Lingdong Shen, Fangxin Shang, Yehui Yang, Xiaoshuang Huang, and Shining Xiang. Segicl: A universal in-context learning framework for enhanced segmentation in medical imaging. *arXiv preprint arXiv:2403.16578*, 2024. 2, 3

[52] Amber L Simpson, Jacob Peoples, John M Creasy, Gabor Fichtinger, Natalie Gangai, Krishna N Keshavamurthy, Andras Lasso, Jinru Shia, Michael I D'Angelica, and Richard KG Do. Preoperative CT and survival data for patients undergoing resection of colorectal liver metastases. *Scientific Data*, 11(1):172, 2024. 4

[53] Karen-Helene Støverud, David Bouget, Andre Pedersen, Håkon Olav Leira, Thomas Langø, and Erlend Fagertun Hofstad. AeroPath: An airway segmentation benchmark dataset with challenging pathology. *arXiv preprint arXiv:2311.01138*, 2023. 4, 7

[54] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3D medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20730–20740, 2022. 1, 3

[55] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1

[56] National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine*, 365(5):395–409, 2011. 4

[57] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1

[58] Constantin Ulrich, Fabian Isensee, Tassilo Wald, Maximilian Zenk, Michael Baumgartner, and Klaus H Maier-Hein. Multi-dataset approach to medical image segmentation: Multitalent. In *BVM Workshop*, pages 78–78. Springer, 2024. 1

[59] Junde Wu and Min Xu. One-prompt to segment all medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11302–11312, 2024. 2, 3

[60] Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023. 2, 3

[61] Ziheng Zhao, Yao Zhang, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. One model to rule them all: Towards universal segmentation for medical images with text prompts. *arXiv preprint arXiv:2312.17183*, 2023. 1, 3, 4