One Model to Rule them All: Towards Universal Segmentation for Medical Images with Text Prompts

Ziheng Zhao[1,2], Yao Zhang[2], Chaoyi Wu[1,2], Xiaoman Zhang[1,2],
Ya Zhang[1,2], Yanfeng Wang[1,2,†] and Weidi Xie[1,2,†]

[1]CMIC, Shanghai Jiao Tong University [2]Shanghai AI Laboratory

Abstract. In this study, we focus on building up a model that can **S**egment **A**nything in medical scenarios, driven by **T**ext prompts, termed as **SAT**. Our main contributions are three folds: (i) on data construction, we combine multiple knowledge sources to construct a multi-modal medical knowledge tree; Then we build up a large-scale segmentation dataset for training, by collecting over 11K 3D medical image scans from 31 segmentation datasets with careful standardization on both visual scans and label space; (ii) on model training, we formulate a universal segmentation model, that can be prompted by inputting medical terminologies in text form. We present a knowledge-enhanced representation learning framework, and a series of strategies for effectively training on the combination of a large number of datasets; (iii) on model evaluation, we train a **SAT-Nano** with only 107M parameters, to segment 31 different segmentation datasets with text prompt, resulting in 362 categories. We thoroughly evaluate the model from three aspects: averaged by body regions, averaged by classes, and average by datasets, demonstrating comparable performance to 36 specialist nnU-Nets, *i.e.*, we train nnU-Net models on each dataset/subset, resulting in 36 nnU-Nets with around 1000M parameters for the 31 datasets. We will release all the codes, and models used in this report, *i.e.*, **SAT-Nano**. Moreover, we will offer **SAT-Ultra** in the near future, which is trained with model of larger size, on more diverse datasets. Webpage URL: https://zhaoziheng.github.io/MedUniSeg.

1 Introduction

Medical image segmentation involves identifying and delineating regions of interest (ROIs) like organs, lesions, and tissues in diverse medical images, which plays a crucial role in numerous clinical applications, such as disease diagnosis, treatment planning, and disease progression tracking. Traditionally, radiologists perform manual segmentation to measure volume, shape, and location in a slice-wise manner, which is both time-consuming and challenging to scale with the growing volume of medical data. Consequently, there is a pressing need for automated and robust medical image segmentation methods in clinical settings, to enhance efficiency and scalability.

Recent advancements in medical image analysis have been marked by a surge of deep learning-based approaches. These developments have yielded a spectrum of segmentation models, each trained for specific tasks [49, 48, 3, 52, 5, 28, 47], often referred to as 'specialist' models. While these models demonstrate impressive segmentation capabilities, their major drawback lies in the narrow specialization. Designed and optimized for distinct ROIs and imaging modalities [24, 25, 31, 51, 55], they often fall short in diverse and dynamic clinical environments where adaptability to new conditions and imaging techniques is essential.

There is a growing interest in developing universal models for medical image segmentation [46, 30], by adapting the pre-trained Segment Anything Model (SAM) [35] models from the computer vision community. However, while transferring to medical scenarios, these models trained on natural images suffer from fundamental limitations: (i) recent models typically perform 2D slice segmentation, that are later fused into 3D volumes through post-processing. This approach overlooks the critical contextual information inherent in 3D radiological imaging; (ii) these models often require point or box inputs as prompts, thus is effectively an interactive segmentation model, still requiring considerable manual efforts for use in practice; (iii) they suffer from significant domain gaps, from image statistics to domain-specific medical knowledge, *e.g.*, 'liver is distinctive for positioning in the upper-right part of the abdominal cavity, filling almost the entire right hypochondrium and often extending into the left hypochondrium as far as the mammillary line'.

In this paper, we present the first knowledge-enhanced universal model for 3D medical volume segmentation,

† *Corresponding author. Email addresses: {Zhao_ Ziheng, weidi}@sjtu.edu.cn*

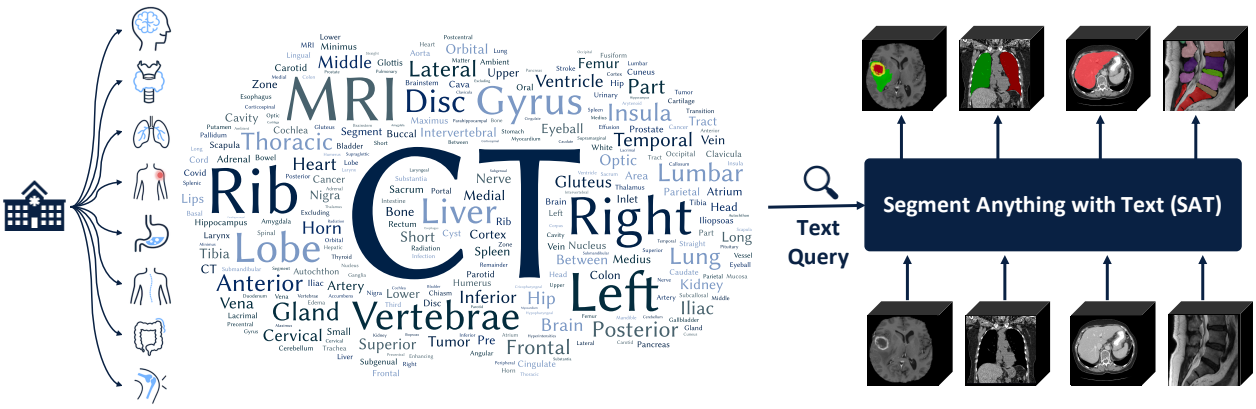

Figure 1 | Segment Anything with Text (SAT) utilizes text prompts as queries and 3D volumes as inputs to perform a wide array of medical image segmentation tasks across different modalities, anatomies, and body regions.

with language/text as prompt, termed as SAT. In practice, our model can effectively take 3D volume as visual inputs along with text prompts, to seamlessly tackle various medical image segmentation tasks, across modalities, anatomies, and body regions, as illustrated in Fig. 1. Specifically, we make the following contributions:

On dataset standardization, we construct a knowledge tree based on multiple medical knowledge source, encompassing thousands of anatomy concepts throughout the human body, accompanied by segmentation/grounding location annotations as visual atlas, and corresponding definitions. As for segmentation training, we curate over 11K 3D medical image scans with 142K anatomical segmentation annotations, covering 362 categories from 31 publicly available medical segmentation datasets. **Note that**, we will continually expand the diversity and volume of training datasets.

On architecture design and training strategy, we propose a universal medical segmentation model that leverages text prompts to guide segmentation, this enables flexible segmentation across a spectrum of medical imaging modalities. Specifically, we adopt knowledge-enhanced representation learning, leveraging textual anatomical knowledge and atlas segmentation of specific anatomical structure to train the text encoder. Through this training process, the visual features of these anatomical structures are aligned with their corresponding text descriptions in the latent space. This alignment facilitates the use of text embeddings as queries in a Transformer-based architecture, enabling the model to selectively extract task-specific features for precise segmentation of the queried target.

On evaluation, we train two models of varying sizes, to satisfy the requirements from different computational resources, namely, **SAT-Nano**, and **SAT-Ultra**. However, note that in this paper, we will exclusively focus on the training details and performance evaluation of SAT-Nano. In the following sections, we term the model SAT for simplicity. We demonstrated that guided by medical domain knowledge, SAT achieves superior performance for universal medical segmentation on 3D inputs. Our comprehensive evaluation covers various perspectives, including region-wise average, organ-wise average, and dataset-wise average. It is worth noting that SAT-Nano with only 107M parameters, can be flexibly used with text prompts for all 31 downstream datasets, showing performance comparable to specialist nnU-Net models trained individually on each dataset, *i.e.*, results from 36 separate nnU-Net models[1].

2 Related work

2.1 Specialist Medical Image Segmentation

Recent years have witnessed significant advancements in the filed of medical image segmentation for diverse clinical applications [54, 7, 47, 58, 3, 6, 49, 33, 48, 28, 45, 64]. Since the success of convolutional neural

[1] When training nnU-Net models, Totalsegmentator is divided into 5 subsets following [64], and CHAOS are divided into MRI subset and CT subset.

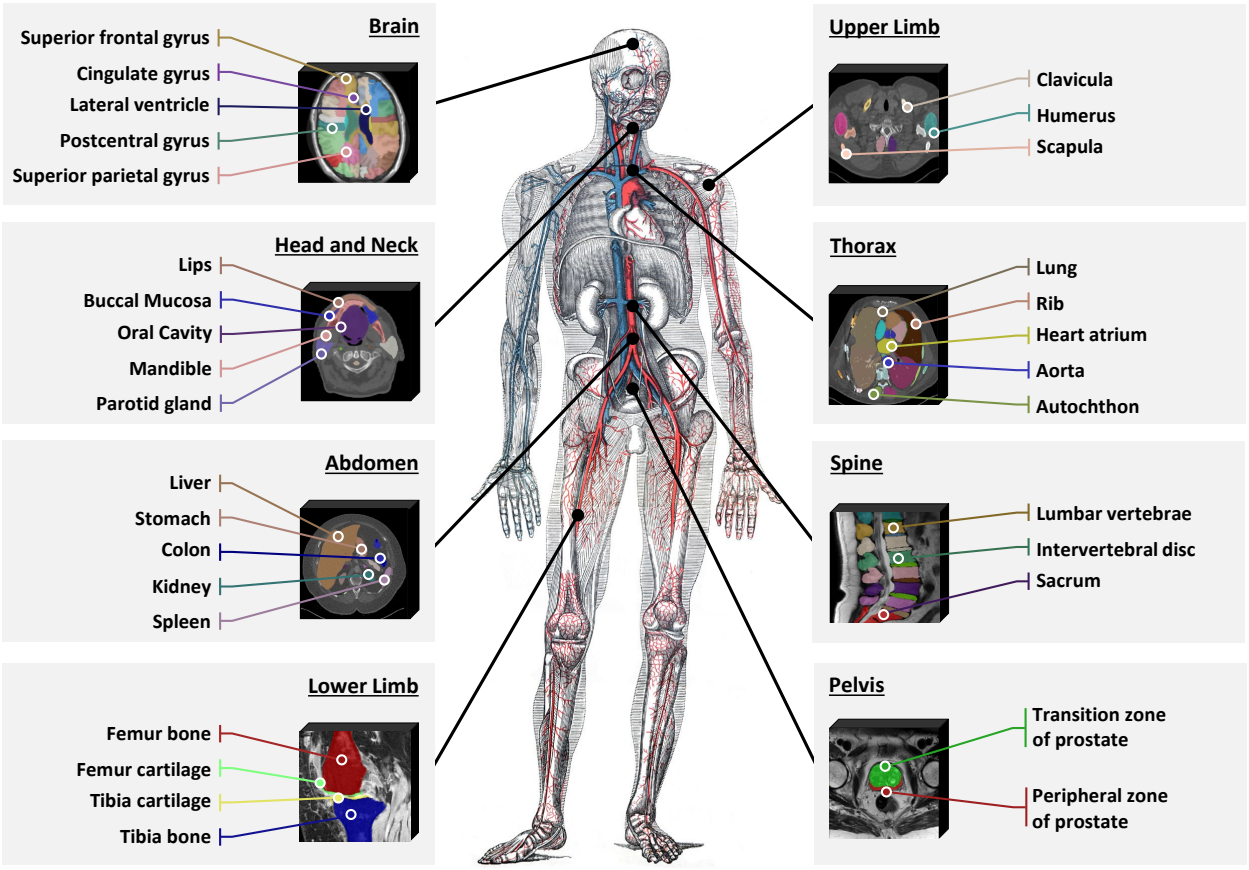

Figure 2 | Overview of our dataset for training universal segmentation model. Specifically, we collect datasets from a wide range of segmentation tasks across various modalities and anatomical regions of the human body, including the brain, head and neck, thorax, spine, abdomen, upper limb, lower limb, and pelvis.

networks (CNNs), U-Net [55, 14, 31] have shown dominant performance for medical image segmentation, and the variants [51, 17, 41, 71, 81, 56, 75] are further developed to dive into specific clinical tasks. Meanwhile, the emerging Transformers [25, 10, 24, 69, 73, 12, 78] step in this field and obtain promising results. This stream of specialist methods typically employ a custom-designed solution and a supervised learning framework. Despite the success in specific segmentation tasks, it requires to be trained from scratch on a new segmentation task, posing limitation on the generalisability. More importantly, the data scarcity for specific tasks often significantly hamper the application in diverse and practical clinical scenarios. In this paper, we present a generalist model that covers a wide spectrum of medical image segmentation tasks.

2.2 Generalized Medical Image Segmentation

In the Medical Segmentation Decathlon [3], nnU-Net [31] takes a promising step towards generalized medical image segmentation. It offers an automatic pipeline to define the optimal network architecture and data pre-processing for each segmentation task. However, the issues of specialist model remain. It has to involve substantial overhead by training from scratch on each task, and the intrinsic correspondence among segmentation tasks is neglected. To effectively benefit a multitude of clinical segmentation tasks, advanced learning schemes have been proposed for generalized medical image segmentation, including multi-task learning [53, 76, 21, 20], transfer learning [11, 16, 63], and continual learning [74, 32, 80]. Multi-task learning incorporates multiple imaging modalities, *e.g.,*multi-parametric MR and multi-phase CT [53, 76], or anatomies, *e.g.,*multi-organ segmentation [20]. In contrast, transfer learning achieves generalized medical image segmentation by transferring segmentation model from one task to another, *e.g.,*from CT to MR image

segmentation and vice versa [11]. Continual learning incrementally involve novel tasks into the segmentation model and avoid forgetting the old ones at the same time [32]. However, these methods requires either a limited pre-defined set of tasks or substantial tuning on novel tasks, and thus can hardly scale up. The proposed method exploits textual prompt learning, *i.e.,*the term of the organs or lesions, to smoothly steer the segmentation model to new tasks without costly training or tuning.

2.3 Universal Medical Image Segmentation

Universal medical image segmentation paves a way to superior generality and scalability by covering a wide range of organs and lesions across various medical imaging modalities and anatomies [30]. The universal model is usually driven by visual or textual prompts to adapt to various segmentation tasks. Since the significant success of SAM [35] in computer vision, several approaches follow the interactive framework and leverage box and/or point prompts to finetune the model on medical images [46, 30, 13, 62, 22]. Other methods also exploit image examples as visual prompts [9] or integrate visual prompts with textual prompts [18]. Despite showing promising results, these semi-automated approaches rely on carefully curated visual prompts, that may face critical challenges in clinical practice. Meanwhile, the typical 2D slice-wise segmentation in most visual prompting methods [46, 30, 13] inevitably leads to suboptimal context encoding for 3D medical imaging [22]. In contrast, pure textual prompts save the effort for burdensome interaction, especially when scaling up to tens of thousands of medical images. Several works have made successful attempts to adapt textual or learnable prompts for abdominal organ and lesion segmentation [43, 70], encouraging us to take a further step towards universal medical image segmentation. However, these studies still require a pre-defined set of categories and thus suffer from the shortage of medical domain knowledge, leading to limited generality across diverse medical segmentation tasks. To bridge this gap, we explicitly leverage multimodal medical knowledge composed of anatomy atlas and definition via visual language pre-training for universal medical image segmentation.

2.4 Knowledge-enhanced Representation Learning in Medical Image Analysis

As accurate medical image analysis relies on expert domain knowledge, several efforts have attempted to involve well-established medical knowledge to strengthen deep learning methods [68, 66]. These methods can be divided into two categories, namely, model-based and learning-based methods. Specifically, the model-based methods design the network architecture based on the clinical knowledge prior [40, 23] or mimicking the clinical practice [15, 19]. The learning-based methods exploit a multi-task learning scheme that takes medical knowledge as either inputs or additional supervision [65, 79]. Additionally, in the recent literature, [67, 72, 39, 77] adopt knowledge-enhanced visual-language representation learning and demonstrate superior performance in downstream classification tasks. Our method aligns with the representation learning methods. While all existing methods mainly focus on incorporation with domain knowledge in text form, we explore multimodal medical knowledge alignment for both text and anatomical atlas, facilitating fine-grained visual-language knowledge injection while training universal segmentation model.

3 Dataset

Toward building a universal segmentation model in medical imaging, with text as prompts, we collect two types of data in this work, namely, the domain knowledge for knowledge-enhanced representation learning, and medical segmentation data. In this section, we will describe them sequentially.

3.1 Domain Knowledge

In this work, we mainly exploit two sources of domain knowledge, namely, e-Anatomy[2] and UMLS[3]. e-Anatomy provides multimodal knowledge in the form of anatomical atlas and their definition, while UMLS refers to a comprehensive knowledge graph, in the form of concept definitions, as well as relations between them. Both knowledge sources are systematic, with anatomies and concepts from all human body regions. Two examples are shown in Fig. 3 (a) and (b).

[2]https://www.imaios.com/en/e-anatomy
[3]https://www.nlm.nih.gov/research/umls

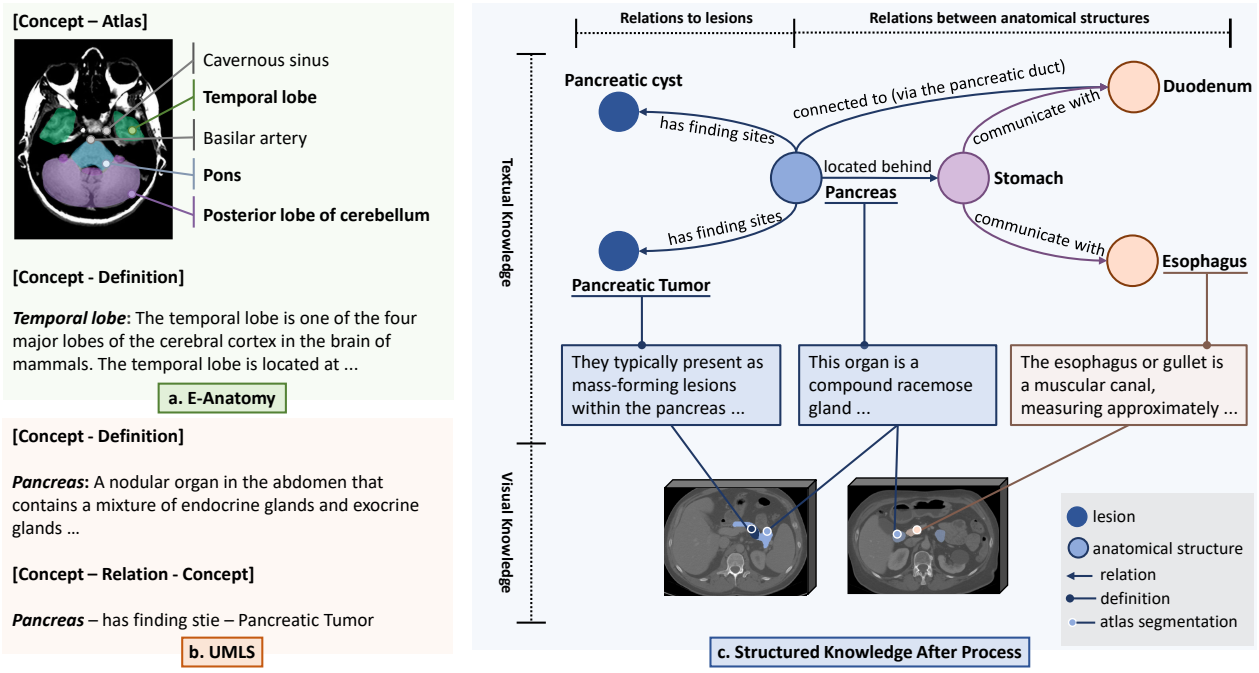

Figure 3 | The medical knowledge used in the visual-language pretraining. (a) e-Anatomy provides an anatomy atlas (including both segmentations and point locations on the image), as well as a detailed definition for each concept. In this example, atlas segmentation is marked with colors. (b) UMLS contains a wide range of concept-definition pairs and tremendous relationships between concepts. (c) By combining all the knowledge sources collected, an implicit medical knowledge tree is constructed. All definitions are partially displayed for conciseness.

e-Anatomy. e-Anatomy contains high-quality anatomy and imaging content atlas, which covers over 8,900 anatomical structures with images in CT, MRI, radiographs, anatomical diagrams, and nuclear images. The anatomical structures are marked on atlas images in the form of segmentations or point locations. Meanwhile, anatomical structures are also equipped with a definition covering their anatomical characteristics, such as their location, shape, and relations to other anatomical structures. We reorganized the image slices into 3D scans and manually filtered out non-radiology scans. We finally derive 321 scans, on which there are 2,224 segmentation and 57,942 point annotations. Note that the same anatomical structure may appear repeatedly, *e.g.*, in different scans of the same body region. On these scans, 6,500 unique anatomical structures and their definitions are collected. As the links between concepts also matter while constructing the implicit knowledge graph, we adopt GPT4 [2] to extract the relations between anatomical structures in their definitions with the following prompt:

> This is the description of xxx. Please help me find its relations with other anatomical structures in radiological images. Summarize them with the template: Relation: xxx (relational preposition), Anatomical structure: xxx (name of another anatomical structure).
> For example, "Relation: situated below, Anatomical structure: xxx", "Relation: connected to (via xxx), Anatomical structure: xxx" ... ...

For instance, this is a sentence in the definition for "pancreas": "Its secretion, the pancreatic juice, carried by the pancreatic duct to the duodenum.". GPT4 extracts the relation: "pancreas, connected to (via the pancreatic duct), duodenum".

UMLS. Unified Medical Language System (UMLS) [8] is a knowledge source of biomedical vocabulary developed by the US National Library of Medicine[4]. It integrates a wide range of concepts from more than 60 families of biomedical vocabularies, each equipped with a Concept Unique Identifier (CUI) and definition. It

[4]https://www.nlm.nih.gov

also contains the relations among these concepts. Following [72], we extract 229,435 biomedical terminologies and definitions, as well as 1,048,575 relationship triplets, implicitly composing a knowledge graph on these terminologies.

Segmentation Datasets. We collect the segmentation datasets that cover 362 anatomical structures and lesions from 8 human body regions. They naturally have the same format as e-Anatomy, *i.e.*, atlas segmentations, we therefore add them as being complementary for multimodal knowledge-enhanced learning. Details will be provided in Sec. 3.2,

In summary, by mixing these data, we construct an implicit multimodal medical knowledge tree on them. As demonstrated in Fig. 3 (c), the concepts (including both anatomical structures and lesions) are linked via the relations and further extended with their definitions, containing their characteristics. Additionally, some are further mapped to segmentations or locations on the atlas images, demonstrating their visual features that may hardly be described purely by text.

3.2 Segmentation Data

To equip our universal segmentation model with the ability to handle segmentation tasks of different targets across various modalities and anatomical regions, we collect and integrate 31 diverse publicly available medical segmentation datasets, totaling 11,462 scans including both CT and MRI and 142,254 segmentation annotations spanning 8 different regions of the human body: Brain, Head and Neck, Upper Limb, Thorax, Abdomen, Plevis, and Lower Limb. The dataset is termed as SAT-DS, and more details are present in Tab. 1. **Note that**, some public datasets are not mutually exclusive, *e.g.*, KITS21 [28] and KITS19 [27], we thus only collect the latest version, to avoid redundancy and potential leaking in train-test split.

Before mixing these datasets for training, two challenges remain: (i) the anatomical targets from each dataset must be integrated into a unified annotation system. The clinic demands beneath each dataset are different, resulting in different annotation standards and granularity. Meanwhile, since most datasets are annotated for training specialist models like nnU-Net [31], precise and consistent terminology or expression for anatomical targets is often ignored. Therefore, a unified label system is demanded to avoid potential contradictions when training on mixed datasets. (ii) some critical image statistics, such as intensity distribution and voxel spacing vary from dataset to dataset, hindering the model from learning consistent image representations across datasets. In the following, we present details for dataset integration and pre-processing, and how we address the abovementioned challenges.

Unifying Label System. While integrating different datasets, three procedures are performed to ensure a unified annotation system: (i) we manually check each anatomical target in each dataset and assign a medical terminology to it, which is guaranteed to be precise and unambiguous across datasets. For instance, the targets that require distinction between orientations, such as the left lung and right lung, are always identified according to the left and right of the human body. And the same anatomical targets from different datasets are named consistently. For example, the i-th lumbar vertebrae in both TotalSegmentator [64] and MRSpineSeg [52] are named with format "lumbar vertebrae i (Li)"; (ii) we adjust the annotations to avoid contradictions between overlapped classes. For example, when lesions are merged into the organs or tissues where they occur, the affected organs and tissues are annotated consistently with the healthy ones; (iii) the same anatomy may have been annotated with different hierarchies in different datasets. In such cases, we manually merge the fine-grained classes to generate additional classes as a complement to close the gap between datasets. For example, sub-regions of liver in Couinaud Liver [59] are merged and added as a new class "liver". As we will keep collecting datasets to increase the scale of SAT-DS, such a label system will be maintained and updated continuously.

Dataset Pre-processing. Since properties of each dataset may greatly impact the training of the segmentation network [31], such as intensity distribution and voxel spacing, we deliberately apply some normalization procedures to all the datasets to ensure uniformity and compatibility between them. *Firstly*, all the images are reoriented to specific axcodes, respaced to a voxel size of $1 \times 1 \times 3\ mm^2$ and cropped to non-zero region. *Secondly*, we apply different intensity normalization strategies to CT images and MRI images, specifically, for CT images, intensity values are truncated to $[-500, 1000]$ and applied z-score normalization. For MRI images, intensity values are clipped by 0.5% and 99.5% of the image, and then z-score normalized. During

Table 1 | The datasets we collect for training the universal segmentation model.

Dataset Name	#Scans	#Classes	#Annotations	Has anomaly	Region
CT Data					
Challenge 4C2021	60	1	60	Yes	Head and Neck
HAN Seg [54]	41	41	1,681	No	Head and Neck
ACDC [7]	300	4	1,200	No	Thorax
COVID-19 CT Seg [47]	20	4	80	Yes	Thorax
LUNA16 [58]	712	4	2,847	No	Thorax
MSD Lung [3]	63	1	63	Yes	Thorax
NSCLC [6]	85	2	162	Yes	Thorax
AbdomenCT1K [49]	988	4	3,950	No	Abdomen
CHAOS (CT) [33]	20	1	20	No	Abdomen
Couinaud [59]	161	10	1,599	No	Abdomen
FLARE22 [48]	50	15	746	No	Abdomen
KITS21 [28]	299	3	746	Yes	Abdomen
MSD Colon [3]	126	1	126	Yes	Abdomen
MSD HepaticVessel [3]	303	2	606	Yes	Abdomen
MSD Liver [3]	131	2	249	Yes	Abdomen
MSD Pancreas [3]	281	2	562	Yes	Abdomen
MSD Spleen [3]	41	1	41	Yes	Abdomen
SLIVER07 [26]	20	1	20	No	Abdomen
WORD [45]	150	18	2,700	No	Abdomen
TotalSegmentor [64]	1,202	128	97,349	No	Whole Body
MRI Data					
Brain Atlas [57]	30	108	3,240	No	Brain
BrainPTM [4]	60	7	408	No	Brain
BraTS2021 [5]	5,004	4	19,680	Yes	Brain
ISLES2022 [29]	500	1	492	Yes	Brain
MSD Hippocampus [3]	260	3	780	No	Brain
WMH [36]	170	1	170	No	Brain
SKI10 [38]	99	4	396	No	Upper Limb
MSD Heart [3]	21	1	20	No	Thorax
CHAOS (MRI) [33]	60	5	300	No	Abdomen
MRSpineSeg [52]	91	23	1,783	No	Spine
MSD Prostate [3]	64	2	124	No	Pelvis
PROMISE12 [42]	50	1	50	No	Pelvis
Summary	11,462	362	142,254	/	/

training, we randomly crop the image patch with a fixed size of $288 \times 288 \times 96$. Random zoom-in, zoom-out, and intensity scaling are applied for data augmentation.

Dataset Statistics. After integrating datasets, we derive a segmentation data collection that covers 362 segmentation classes, far outpacing each single dataset in both diversity and scale. We divide human body

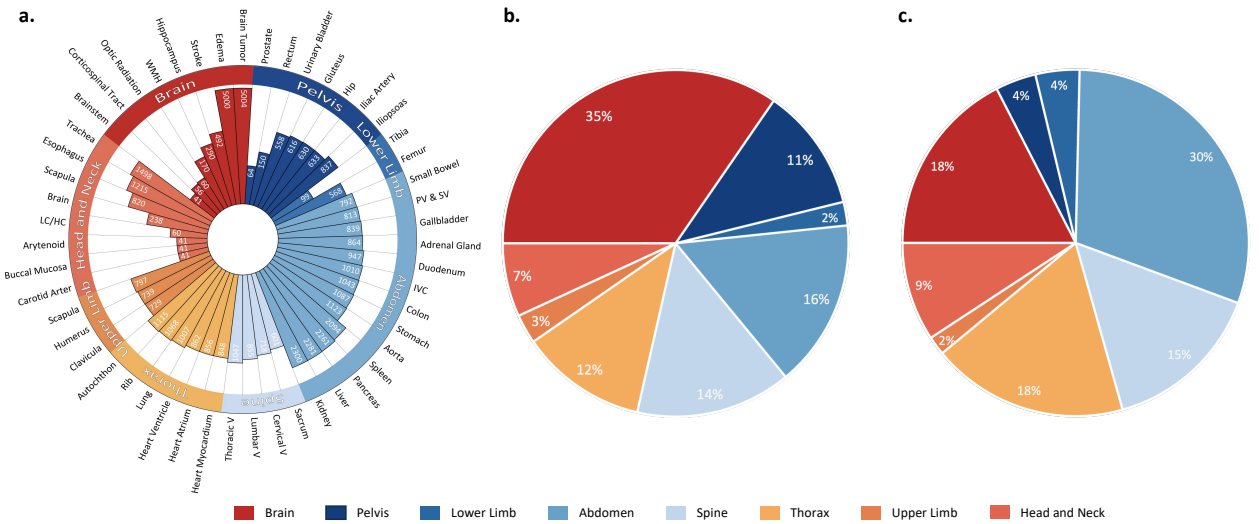

Figure 4 | Statistics of SAT-DS across different anatomical regions. (a) Annotation num of some representative classes in each anatomical region; (b) Number of classes in each anatomical region; (c) Number of annotations in each anatomical region. LC/HC: laryngeal/hypopharyngeal cancer, WHM: white matter hyperintensities, PV&SV: portal vein and splenic vein, IVC: inferior vena cava, Thoracic V: thoracic vertebrae, Cervical V: cervical vertebrae, Lumbar V: lumbar vertebrae.

into eight regions, and classify each class into them manually. Fig. 4 (a) and (b) show the distribution of classes and annotations across different human body regions. We further show the distribution of some example classes in each region in Supplementary Fig. 7. The extensive range of categories and regions lays the foundation for the SAT's wide application scenarios. The data collection is more than **twice** the size of the largest dataset (BraTS2021) in terms of volume number, and nearly **triple** the most comprehensive dataset (TotalSegmentator) in terms of the class number. **Note that**, we will continually expand the diversity and volume of training datasets.

Discussion. In the process of building SAT-DS, we merge a wide range of segmentation tasks, and establishing a unified label system by using natural language/text. Generally speaking, there are two advantages to do this: (i) natural language is powerful and discriminative, that enables better differentiation of the medical terminologies in the language embedding space; (ii) as shown in previous work [67, 72, 39, 77], knowledge-enhanced representation learning for the text encoder demonstrating promising performance, allowing to learn the implicit or explicit relationships between these segmentation targets. For example, segmenting a specific lobe of the liver requires the exact segmentation of liver as an organ in abdomen cavity, and shall be facilitated refering to other parts of the liver. Therefore, establishing such connections via systematic medical knowledge shall be beneficial.

4 Method

Towards building up our universal segmentation model by text prompt, *i.e.*, SAT, we consider two main stages, namely, multimodal knowledge injection and universal segmentation training. In the following, we start by defining the problem scenario (Sec. 4.1). Then, we show how to structure multimodal medical knowledge and inject it into a text encoder (Sec. 4.2). Lastly, we employ the text encoder to guide our universal segmentation model training on SAT-DS dataset (Sec. 4.3).

4.1 Problem Formulation

Assuming we have a segmentation dataset collection, *i.e.*, $\mathcal{D} = \{(x_1, y_1; T_1), ..., (x_K, y_K; T_K)\}$, where $x_i \in \mathbb{R}^{H\times W\times D\times C}$ denotes the image scan, $y_i \in \mathbb{R}^{H\times W\times D\times M}$ is the binary segmentation annotations of the anatomical targets in the image, and $T_i = \{t_1, t_2, ..., t_M\}$ denote the corresponding medical terminology set.

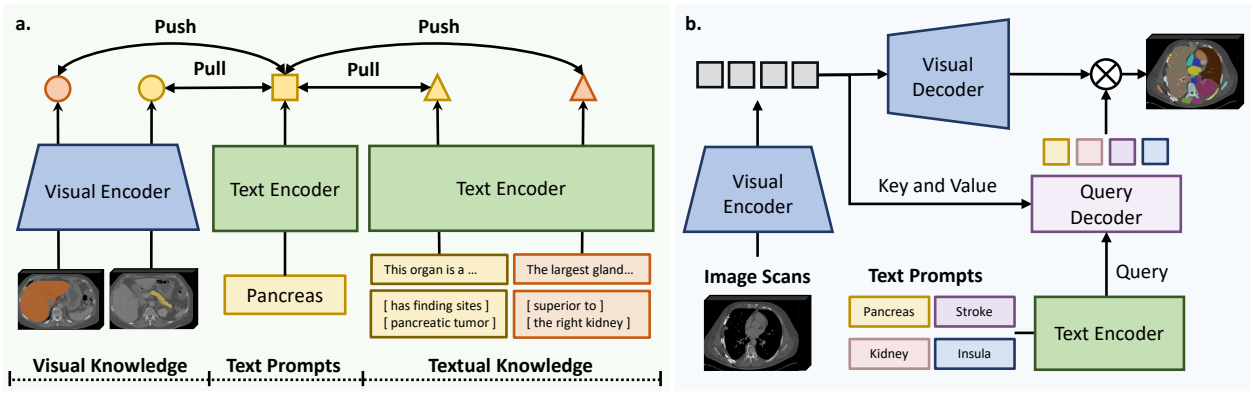

Figure 5 | Overview of SAT. (a) We inject multi-modal medical knowledge into knowledge encoders via contrastive learning. The knowledge is in different formats: atlas segmentation, concepts(terminologies), definitions, and relationships between concepts. We devise different knowledge encoders to embed them; (b) We train a universal segmentation network based on the text prompts from the pre-trained knowledge encoder. It is capable of segmenting a wide range of targets for image scans from different modalities and anatomical regions.

The universal segmentation task can be formulated as:

$$\hat{y}_i = \Phi_{\text{SAT}}(\Phi_{\text{visual}}(x_i), \Phi_{\text{text}}(T_i)), \tag{1}$$

where Φ_{visual} is a visual encoder, Φ_{text} is a text encoder, Φ_{SAT} is a universal segmentation model and ideally, x_i can be an image scan from any modality of the anatomical region, T_i can contain any text-wise medical terminology of interest.

4.2 Multimodal Knowledge Injection

Here, we aim to inject rich multimodal medical domain knowledge into the visual and text encoders. We will start from the procedure for structuring the multimodal medical knowledge data, and further present details to use them for visual-language pre-training, injecting knowledge into the two encoders.

Multimodal Domain Knowledge. As shown in Fig. 5 (a), the data from e-Anatomy, UMLS, and segmentation datasets can be aggregated into two formats:

- **Textual medical concept pair.** Within both e-Anatomy and UMLS, each concept t_i is associated with a human-revised professional definition p_i, constructing pairs of text $(t_i; p_i)$. We also derive a knowledge graph that connects the medical concepts through abundant triple relationships (t_i, r_{ij}, t_j). This graph can be alternatively seen as a specialized text pair, $(t_i + r_{ij}; t_j)$ or $(t_i; r_{ij} + t_j)$, where $+$ means string concatenation. In this way, we can thus unify the two kinds of textual knowledge.
- **Visual medical concept pair.** To align with the segmentation task, we gather pairs consisting of a concept (can be either an anatomical structure or lesion) and its atlas segmentation on the image. Note that, multiple pairs could be extracted from a single image. These pairs share a similar format to the segmentation data, denoted as $(x_i, y_i; t_i)$. e-Anatomy also contains numerous pairs of concepts and approximate grounding locations (points) in the image. For simplicity, we treat this knowledge as a specialized atlas segmentation, involving annotations for only one pixel. This approach unifies the format of the two kinds of visual knowledge.

In summary, all the knowledge can either be represented as pure text description, *e.g.*, t_i, d_i, $t_i + r_{ij}$, $r_{ij} + t_j$, or atlas segmentation (x_i, y_i), and paired further.

Visual-Language Pre-training. As shown in Fig. 5 (a), for the textual concepts, we adopt the text encoder

Φ_{text} to encode them. Specifically, we adopt a standard BERT [34] pre-trained on PubMed abstract [37]:

$$z = \Phi_{\text{text}}(\mathbf{t}),\ \mathbf{t} \in [t_i, d_i, t_i + r_{ij}, r_{ij} + t_j],\ z \in \mathbb{R}^d, \tag{2}$$

where d refers to the feature dimension. For the visual concepts, we adopt the visual encoder Φ_{visual}. Given the excellent robustness and performance of U-Net [31, 55], we apply a standard 3D U-Net encoder to extract image embeddings, further average ROI pooling on the spatial feature map output, based on the atlas segmentation mask. The pooled feature can thus be treated as a representation of the anatomical target on this image.

$$z = \mathcal{F}_{\text{pooling}}(\Phi_{\text{enc}}(x_i); y_i),\ z \in \mathbb{R}^d. \tag{3}$$

Then, we train the two encoders by maximizing the similarities between all positive knowledge pairs, linked by text prompts (medical terminology names), as shown in Fig. 5. Specifically, given $(x_i, y_i; t_i), (t_i; p_i), (t_i + r_{ij}; t_j), (t_i; r_{ij} + t_j)$, for conciseness, regardless their modality format, we all denote the encoded features as z, and for a batch of N pairs $\{(z_1, z_1'), ...(z_N, z_N')\}$, we have:

$$\mathcal{L}_{\text{knowledge}} = -\frac{1}{N}\sum_{i=1}^{N}(\log \frac{\exp(z_i \cdot z_i'/\tau)}{\sum_{k=1}^{N} \mathbb{1}_{i\neq k} \exp(z_i \cdot z_k'/\tau)} + \log \frac{\exp(z_i \cdot z_i'/\tau)}{\sum_{k=1}^{N} \mathbb{1}_{i\neq k} \exp(z_k \cdot z_i'/\tau)}), \tag{4}$$

which is a typical contrastive learning pipeline, with $\tau = 0.07$ as temperature.

Discussion. At a high level, by maximizing the similarities between positive textual and visual feature pairs, we force the text encoders to construct neural representations for medical concepts from two aspects: (i) through the well-established knowledge graph in text form, the text encoder enables to encode relationships between concepts in the latent space; (ii) the model captures the characteristics of anatomical structures and lesions via both visual atlas segmentations and detailed definitions. Therefore, in contrast to the one-hot labeling that treats each anatomical target as of being independent, such continuous neural representation shall provide more helpful guidance for the segmentation task.

4.3 Universal Segmentation Training

With the pre-trained visual and text encoders, we now continue the procedure for building the universal segmentation model with text as prompts. Fig. 5 (b) demonstrates the overall framework. Specifically, apart from the pre-trained visual and text encoders, the segmentation model consists of three more components: a visual decoder Φ_{dec}, a query decoder Φ_{query}, and a mask generator.

Given one sample in the segmentation dataset collection $(x_i, y_i; T_i)$, for conciseness, we next will simplify the terminology class set T_i as a certain terminology element t_i in it, and by performing the following procedure on each element in T_i, we will get the complete segmentation results.

Terminology Encoding. We use the pre-trained text encoder to generate neural embedding for any anatomical target terminology as a text prompt for segmentation:

$$z_i = \Phi_{\text{text}}(t_i),\ z_i \in \mathbb{R}^d. \tag{5}$$

Note that, after pre-training the text encoder with domain knowledge injection, z_i should contain both the textual background information and visual information from atlas samples.

Visual Encoding. We adopt the pre-trained visual encoder as explained in Sec. 4.2, and continue training it:

$$V_i = \{v_{i1}, v_{i2}, ..., v_{iS}\} = \Phi_{\text{visual}}(x_i),\ v_s \in \mathbb{R}^{H_s \times W_s \times D_s \times d}, \tag{6}$$

where V_i is the multi-scale feature maps from U-Net encoder layers, and H_s, W_s, D_s are the spatial resolutions at different layers.

Visual Decoding. In the visual decoder, the feature maps from the encoder are gradually upsampled with

skip connections, effectively following the U-Net architecture [31, 55], ending up with per-pixel dense feature:

$$u_i = \Phi_{\text{dec}}(V_i),\ u_i \in \mathbb{R}^{H\times W\times D\times d'}, \tag{7}$$

where d' is the dimension for the per-pixel dense feature after recovering to the original resolution.

Query Decoder. Although a general representation of the anatomical target is derived from the pre-trained text encoder with a text prompt, visual variations may remain from patient to patient, we thus insert a transformer-based query module to further enhance the text prompts with visual clues. In practice, this module consists of 6 standard multi-head cross attention layers [61] that treat text prompt embedding as query and the high-level latent visual embedding from the U-Net encoder as key, values, formulated as:

$$q_i = \Phi_{\text{query}}(V_i, z_i),\ z_i \in \mathbb{R}^d. \tag{8}$$

Therefore q_i can be seen as an adapted representation of the anatomical target in specific image scan x_i.

Mask Generator. After conducting pixel-wise dot-product between the representation of the anatomical target and the fine-grained per-pixel embedding, we can acquire a per-pixel segmentation:

$$\hat{y}_i = g(q_i)\cdot u_i,\ \hat{y}_i \in \mathbb{R}^{H\times W\times D}, \tag{9}$$

where $g(\cdot)$ is a feed-forward layer projecting q_i to a consistent dimension with the dense feature map u_i.

Training Objective. Following [31], we adopt a loss function as the sum of binary cross-entropy loss and dice loss. For a sample with M classes and C voxels, we denote $p_{\text{c,m}}$ and $s_{\text{c,m}}$ the prediction and ground-truth for c-th pixel respectively on class m, the loss is:

$$\mathcal{L} = \underbrace{-\frac{1}{M}\sum_{\text{m}=1}^{M}\frac{1}{C}\sum_{\text{c}=1}^{C} p_{\text{c,m}}\cdot\log s_{\text{c,m}}}_{\text{Binary Cross Entropy Loss}} + \underbrace{(1-\frac{2\sum_{\text{i}=1}^{M}\sum_{\text{c}=1}^{C} p_{\text{c,m}}\cdot s_{\text{c,m}}}{\sum_{\text{m}=1}^{M}\sum_{\text{c}=1}^{C} p_{\text{c,m}}^2+\sum_{\text{m}=1}^{M}\sum_{\text{c}=1}^{C} s_{\text{c,m}}^2})}_{\text{Dice Loss}} \tag{10}$$

5 Training Strategies

We encounter several challenges in training on a combination of a large number of heterogeneous medical datasets in 3D format. In this section, we provide details on the training strategies we used.

Two Stage Visual-Language Training. The visual-language training procedure introduced in Sec. 4.2 is divided into two stages. At the first stage, the Biomed BERT [37] is finetuned on the text knowledge via contrastive learning, *i.e.,* the textual medical concept pairs. The maximal sequence length after tokenization is 256, for even longer text input, we apply random truncation to fully exploit the knowledge in the long text. At the second stage, to align text representations and visual features, we apply contrastive learning between the finetuned text encoder and the U-Net encoder on the visual medical concept pair. The U-Net encoder will be used as an initialization later for the visual encoder in SAT. Note that, we freeze the text encoder in the second stage to avoid knowledge forgetting.

Pre-processing Segmentation Dataset. We empirically take voxel spacing as $1\times 1\times 3mm^2$ and patch size $288\times 288\times 96$, based on two empirical considerations: (i) when mixing the various datasets, scans with a wide range of voxel spacings ought to be normalized before processing in the same convolutional network. While resampling to larger voxel spacing may lose information, smaller voxel spacing will generate artifacts, (ii) a larger receptive field generally ensures better segmentation performance for most targets, however, increasing patch size will result in higher computational cost. We strike a balance between them based on the computational resources in use.

Visual Backbone Selection. Despite the recently transformer-based architectures [25, 24, 43] have gained much attention, convolution-based architectures [55, 31] remains to be mainstream in medical segmentation tasks, and are recently proven superior in multi-dataset training [60]. Combining with our initial trial and

error on these two architectures, we adopt the standard U-Net architecture for its robustness, computation efficiency and versatility.

Balancing Segmentation Datasets. To balance between all datasets, we set the sampling strategy for each scan based on two intuitions respectively: (i) training case number varies significantly from dataset to dataset, which should be alleviated. We follow [60] and set the sampling weight of all scans in a dataset as the inverse proportion to $\sqrt{N}$, N is the number of training cases in the dataset; (ii) scans with larger spatial resolution or more annotated classes should be sampled more as they are often harder to learn. Thus, we repeat such scans for R times in the sampling pool, where $R = \frac{H \times W \times D}{288 \times 288 \times 96} \times \frac{M}{32}$. H, W, D are the spatial resolution of the scan, c is the number of annotated classes on it, and 32 is the maximal number of text prompts in a batch.

Balancing Segmentation Classes. While cropping the image scan, it's a common practice [31] to oversample foreground crops, *i.e.*, crops containing at least one segmentation target. However, weighting these crops evenly may ignore the unbalanced spatial distribution of segmentation targets. For example, in large scans with numerous annotations, tiny targets are harder to sample and thus may be ignored by the model. Thus, in foreground oversampling, we give more weight on regions with more segmentation targets.

Curriculum Learning on Segmentation. Within the segmentation targets involved in training, there exist some anatomical hierarchies. For example, learning to segment the liver should ideally help the segmentation of the lobes of the liver. Therefore, we take a curriculum learning strategy, *i.e.*, gradually adding segmentation datasets into training. Datasets with fine grained classes such as Brain Atlas [57] and Couinaud Liver [59] are added in the last stage of training.

Other Hyperparameter Details. For **SAT-nano**, we adopt a U-Net with 6 blocks in-depth, each block has 2 convolutional layers and 3×3 size each kernel. This is consistent with the U-Net encoder in visual-language pretraining. The query decoder is a 6-layer standard transformer decoder with 8 heads in each attention module. Feature dimensions of a text prompt $d = 768$ and per-pixel embedding $d' = 64$. For images with multiple segmentation targets, we set the maximal text prompts sampled in a batch of up to 32. A combination of cross-entropy loss and dice loss is applied as supervision at training time. We use AdamW [44] as the optimizer with cosine annealing schedule, maximal $lr = 1 \times 10^{-4}$, 10000 steps for warm-up. The model is trained on 8 NVIDIA A100 GPUs with 80GB memory, using maximal batch size.

6 Experiment

6.1 Evaluation Datasets

We evaluate **SAT-nano** on all the 31 segmentation datasets in the collection. As there is no existing benchmark for evaluating the universal segmentation model, we randomly split each dataset in the collection into 80% for training and 20% for testing with two exceptional cases: (i) datasets may share the same images but with different classes. For example, Couinaud Liver provides fine-grained liver segmentation on a subset of MSD Hepatic Vessel. We carefully split the Couinaud Liver to make sure the test set will not be leaked in the train set of MSD Hepatic Vessel; (ii) scans of the same patient but different modalities are treated as different samples when training and evaluating. For example, BraTS2021 contains T1, T1CE, T2, FLAIR scans of each patient. However, they share the same structure on the image. To avoid potential data leaking, we carefully split such datasets by patient id. **We will release our dataset splits to the research community for future model comparison.**

6.2 Baseline

We take nnU-Net [31] as a strong baseline for comparison, which specializes in a series of configurations for each dataset. For a comprehensive evaluation, we train one nnU-Net model on each of the datasets. Note that, following [64], we split Totalsegmentator into 5 subsets and train 5 different nnU-Net models on each of them. Similarly, CHAOS with both MRI and CT images are split into two subsets. When training nnU-Nets on each dataset, we adopt a multi-class segmentation setting and deliver the masks of all categories in this dataset at once. We derive the optimal network architecture and pre-processing pipeline following the default setting of nnU-Net. The detailed network design is shown in Table 7. In summary, we train an assemble of

36 nnU-Nets, that are customized on each dataset and serve as the competitive baseline for evaluation. We adopt the latest official implementation of nnU-Net v2[5] in practice.

6.3 Evaluation Protocols

Given our goal is to develop a universal medical segmentation model, this provides opportunities to evaluate novel perspectives in addition to the traditional evaluation per dataset. Specifically, we conduct the evaluations from three dimensions:

- **Region-wise Evaluation.** In general, anatomical structures from the same human body region are closely connected and more likely to be involved in diagnosis within the same hospital department. Here, we consider the region-wise evaluation, that is to merge all segmentation results of the same region, as to indicate the general performance in this region.
- **Class-wise Evaluation.** As SAT is capable of segmenting a wide range of anatomical targets across the human body, we merge the results from the same classes across datasets to indicate the performance on each anatomical target. Note that, the same class from different modalities is merged as well, *e.g.,* liver in both CT and MRI images.
- **Dataset-wise Evaluation.** Results of the classes within the same dataset are averaged to indicate the performance on this dataset. This is the same as the conventional evaluation protocol of specialist segmentation models trained on a single dataset, and we believe this is less valuable for evaluating universal segmentation models.

Note that, class-wise and region-wise evaluations require averaging results across different nnU-Nets.

6.4 Evaluation Metrics

We quantitatively evaluate the segmentation performance from the perspective of region and boundary metrics [50], *e.g.*, Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD) respectively.

DSC. Dice Similarity Coefficient (DSC) is a standard region-based metric for medical image segmentation evaluation. It measures the overlap between model's prediction P and ground truth G, formally defined as:

$$DSC(P,G) = \frac{2|P \bigcap G|}{|P| + |G|}. \tag{11}$$

NSD. Normalized Surface Distance (NSD) [1] is a boundary-based metric that measures the consistency at the boundary area of the model's prediction P and ground truth G, which is defined as:

$$NSD(P,G) = \frac{|\partial P \bigcap B_{\partial G}| + |\partial G \bigcap B_{\partial P}|}{|\partial P| + |\partial G|}, \tag{12}$$

where $B_{\partial P} = \{x \in \mathbf{R}^3 | \exists \hat{x} \in \partial P, ||x - \hat{x}|| \leq \tau\}$ and $B_{\partial G} = \{x \in \mathbf{R}^3 | \exists \hat{x} \in \partial G, ||x - \hat{x}|| \leq \tau\}$ are the boundary area of model's prediction and ground truth at a tolerance τ, respectively. We set τ as 1 in the experiments.

7 Results

To thoroughly evaluate our universal model **SAT-Nano**, we first compare it to the powerful specialist model, *i.e.*, nnU-Nets trained on each dataset. Driven by the knowledge-enhanced textual prompts, our proposed universal segmentation model is capable of covering a wide spectrum of medical image segmentation tasks, leading to 362 anatomical targets on 8 regions and lesions of the human body across 31 datasets. To comprehensively understand the behavior of **SAT-Nano**, we make an evaluation from the perspective of anatomical regions, classes, and datasets.

Region-wise evaluation. Fig. 6 (a) and (b) present the segmentation performance on 8 regions of human body, including brain, head and neck, thorax, abdomen, pelvis, spine, upper limb, and lower limb, as well as lesions, in terms of DSC and NSD, respectively. Overall, **SAT-Nano** obtains competitive results with

[5] https://github.com/MIC-DKFZ/nnUNet/

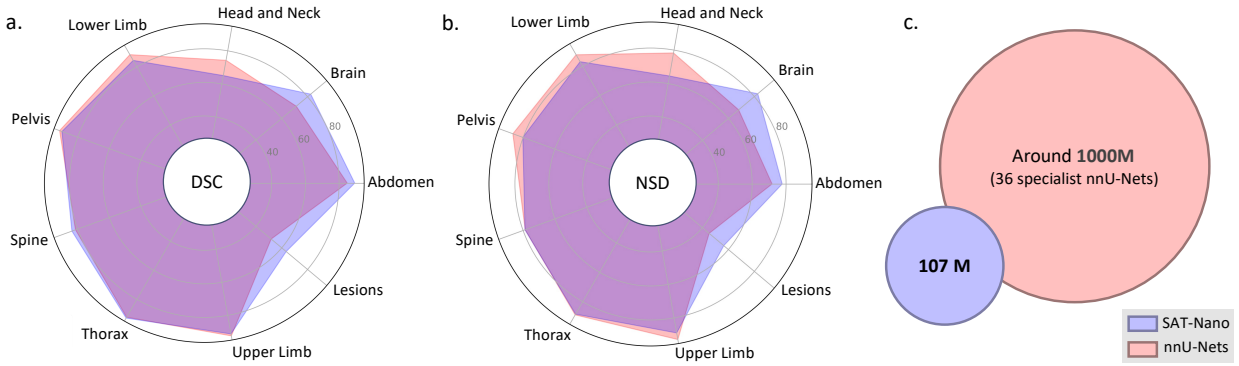

Figure 6 | Comparison between **SAT-Nano** and **nnU-Nets** on different regions and lesions. (a) Dice score comparison; (b) NSD score comparison; (c) Model size comparison. **SAT-Nano** is approximately 10 times smaller than the total size of all the specialist nnU-Nets.

a assemble of 36 specialist nnU-Nets. The detailed results are show in Table 2. Notably, **SAT-Nano** even consistently outperforms nnU-Nets in 4 out of 8 regions, including head and neck, pelvis, upper limb and lower limb. As shown in Fig. 6 (c), apart from the promising segmentation performance, **SAT-Nano** is of significantly small model size of 107M, approximately 1/10 of the assemble of nnU-Nets. Benefit from flexible and robust text prompts, **SAT-Nano** demonstrates promising application in various clinical practice.

Table 2 | Segmentation results on 8 regions of human body and lesion. H&N: head and neck, LL: lower limb, UL: upper limb.

Metric	Method	Abdomen	Brain	H&N	LL	Pelvis	Spine	Thorax	UL	Lesion
DSC↑	nnUNet	89.30	82.45	65.04	84.40	89.94	83.61	92.83	91.22	63.44
	SAT-Nano	84.84	71.26	74.29	88.35	91.41	81.36	92.20	92.40	51.82
NSD↑	nnUNet	77.73	82.64	64.53	83.09	80.19	78.83	88.42	88.98	56.43
	SAT-Nano	71.76	67.91	78.29	87.79	86.44	78.70	88.92	92.92	45.19

Class-wise evaluation. From Table 3 and Table 4, it can be observed that **SAT-Nano** demonstrates promising results on major organs, *e.g.,*brain, heart atrium and ventricle, liver, kidney, prancreas, spleen, stomach, gallbladder, intestine, urinary bladder, and adrenal gland, anatomies, *e.g.,*aorta, inferior vena cava, gluteus, hip, iliopsoas, eyeballs, and skeleton, *e.g.,*clavicula, humerus, rib, vertebrae, sacrum, femur, and tibia. The detailed results of 10 kinds of lesions across whole body are shown in Table 6. **SAT-Nano** still shows promising results, especially on dominant targets with major samples, such as brain tumor.

Dataset-wise evaluation. As shown in Table 5, **SAT-Nano** obtains comparable or even better results on AbdomenCT1K, CHAOS CT, HAN Seg, LUNA16, MRSineSeg, MSD Hippocampus, MSD Liver, MSD Prostate, MSD Spleen, SLIVER07, TotalSegmentator, and WORD. These datasets typically involve major annotations and classes, covering major anatomical targets, as shown in Table 1 and Table 5. Combining with multimodal medical knowledge injection, **SAT-Nano** effectively leverage the intrinsic connections within diverse visual semantics via textual prompts. It demonstrates the potential scaling up **SAT-Nano** with more annotations and classes towards universal segmentation of medical images.

8 Conclusion

In this paper, we promote the progress of medical universal segmentation with text as prompt. In contrast to existing segmentation models that are limited in 2D input and box/point prompts, our model, **SAT**, enables text-driven universal segmentation on various radiologic modalities across 8 main human body parts, greatly reducing the usage difficulty of medical universal segmentation model in practice. In this report, we will release the **SAT-Nano**, which only contains 107M paramters, while demonstrating comparable performance to 36 specialist nnU-Nets, with around 1000M parameters on 31 datasets for segmenting 362 categories.

Table 3 | Segmentation results on typical anatomical structures in abdomen, brain, and pelvis. The results of nnUNet are aggregated from the assemble of 36 specialist models trained separately on the corresponding datasets.

Region	Modality	Anatomical Target	DSC↑		NSD↑	
			nnUNet	SAT-Nano	nnUNet	SAT-Nano
Abdomen	CT	Adrenal gland	87.50	81.65	87.73	88.71
	CT	Aorta	95.73	93.54	88.60	93.34
	CT	Colon	89.59	87.21	75.84	76.46
	CT	Duodenum	85.29	80.93	73.31	73.30
	CT	Gallbladder	86.79	81.27	80.79	79.09
	CT	Inferior vena cava	90.74	90.79	81.07	88.16
	CT	Intestine	89.41	88.02	81.87	75.68
	CT	Kidney	95.71	95.48	91.09	92.29
	CT	Liver	97.71	96.79	88.00	85.71
	CT	Pancreas	88.93	87.94	77.48	79.89
	CT	Small bowel	89.05	86.01	76.74	80.10
	CT	Spleen	96.09	95.26	92.22	91.15
	CT	Stomach	93.28	91.90	81.29	81.03
	MR	Kidney	91.48	88.13	68.50	57.15
	MR	Liver	86.50	83.41	57.69	42.81
	MR	Spleen	83.14	64.23	58.27	35.33
Brain	CT	Brainstem	83.01	77.38	62.30	54.25
	CT	Optic chiasm	54.54	36.25	73.22	59.81
	CT	Pituitary gland	67.02	65.18	76.79	82.87
	MR	Amygdala	83.90	73.99	93.24	83.04
	MR	Basal ganglia	90.81	79.29	97.32	79.93
	MR	Brainstem	95.14	90.41	96.18	85.57
	MR	Cerebellum	96.99	92.30	96.65	78.15
	MR	Cingulate gyrus	86.90	75.14	85.58	57.04
	MR	Corpus callosum	89.30	78.75	94.22	78.84
	MR	Corticospinal tract	62.51	56.33	31.08	45.52
	MR	Frontal lobe	96.15	82.91	92.12	48.61
	MR	Hippocampus	90.13	81.46	98.28	89.83
	MR	Insula	89.60	75.01	92.35	65.53
	MR	Lateral ventricle	89.14	75.61	98.11	79.31
	MR	Occipital lobe	91.33	84.01	79.77	56.57
	MR	Optic radiation	65.22	53.17	35.36	45.70
	MR	Parietal lobe	88.64	83.00	77.30	49.51
	MR	Temporal lobe	92.93	83.94	86.79	62.72
	MR	Thalamus	91.19	84.25	93.32	81.37
	MR	Third ventricle	84.33	75.82	96.71	93.04
	MR	White matter hyperintensities	77.02	64.24	88.88	78.10
Pelvis	CT	Bladder	93.40	91.16	79.87	74.22
	CT	Gluteus maximus	96.85	95.85	85.42	92.28
	CT	Gluteus medius	95.61	95.08	82.25	91.72
	CT	Gluteus minimus	95.63	95.27	85.87	94.79
	CT	Hip	94.62	95.29	92.29	95.10
	CT	Iliopsoas	92.18	91.69	81.22	91.04
	CT	Rectum	80.35	81.42	67.57	66.02
	CT	Urinary bladder	94.70	93.00	79.00	84.68
	MR	Prostate	88.86	82.31	71.90	54.60

Table 4 | Segmentation results on common anatomical structures in head and neck, lower limb, spine, and upper limb. The results of nnUNet are aggregated from the assemble of 36 specialist models trained separately on the corresponding datasets.

Region	Modality	Anatomical Target	DSC↑		NSD↑	
			nnUNet	SAT-Nano	nnUNet	SAT-Nano
Head and neck	CT	Arytenoid	63.03	48.69	79.15	74.51
	CT	Brain	88.81	97.24	84.59	93.89
	CT	Buccal mucosa	59.88	56.20	50.69	51.27
	CT	Carotid artery	81.79	65.54	90.55	77.82
	CT	Cervical esophagus	62.00	63.13	57.53	65.16
	CT	Cochlea	66.89	62.18	85.57	87.96
	CT	Cricopharyngeal inlet	68.54	62.68	68.03	69.77
	CT	Esophagus	90.98	84.53	86.77	85.18
	CT	Eyeball	93.69	92.55	92.66	93.33
	CT	Lacrimal gland	53.46	51.25	64.68	69.93
	CT	Larynx - glottis	77.29	72.77	82.93	83.46
	CT	Larynx - supraglottic	81.16	79.25	76.65	79.33
	CT	Lips	67.26	61.07	55.74	55.71
	CT	Mandible	94.57	91.89	96.36	93.62
	CT	Optic nerve	64.86	61.82	80.91	82.48
	CT	Oral cavity	88.30	86.45	59.12	60.83
	CT	Parotid gland	81.77	79.93	61.31	64.04
	CT	Scapula	96.18	93.92	94.53	95.70
	CT	Submandibular gland	82.75	78.88	74.29	74.64
	CT	Thyroid	90.39	86.34	92.81	90.63
	CT	Trachea	95.14	92.45	93.28	92.54
Lower limb	CT	Femur	92.76	94.69	88.76	93.24
	CT	Head of femur	92.78	89.95	84.97	80.40
	MR	Femur bone	98.25	94.87	96.80	86.16
	MR	Femur cartilage	78.87	68.39	91.39	88.18
	MR	Tibia bone	98.19	95.50	97.11	90.26
	MR	Tibia cartilage	77.30	66.75	92.21	88.31
Spine	CT	Cervical vertebrae	95.50	88.66	94.80	89.74
	CT	Lumbar vertebrae	95.94	93.55	95.14	93.94
	CT	Sacrum	94.92	95.41	92.13	94.84
	CT	Thoracic vertebrae	97.79	95.67	97.22	96.58
	MR	Intervertebral discs	74.72	85.99	63.30	86.94
	MR	Lumbar vertebrae	71.89	82.15	56.70	69.85
	MR	Sacrum	67.31	78.34	55.61	67.32
	MR	Thoracic vertebrae	55.70	76.86	45.67	64.92
	MR	Vertebrae	69.23	81.98	54.92	60.13
Thorax	CT	Autochthon	94.98	93.88	78.49	90.35
	CT	Heart atrium	95.00	94.17	81.59	90.17
	CT	Heart myocardium	93.92	92.01	82.75	90.43
	CT	Heart ventricle	95.46	93.48	80.12	88.05
	CT	Lung	99.21	95.08	97.36	82.08
	CT	Pulmonary artery	94.12	93.01	83.98	90.23
	CT	Rib	94.06	91.89	93.91	94.66
	CT	Thoracic cavity	70.58	95.60	52.46	74.71
	MR	Myocardium	92.84	79.40	97.68	68.46
	MR	Ventricle cavity	91.44	86.74	97.89	66.32
Upper limb	CT	Clavicula	96.03	94.96	94.29	95.87
	CT	Humerus	87.96	86.26	84.41	85.46

Table 5 | Segmentation results on 31 individual datasets. The results of nnUNet are aggregated from the assemble of 31 specialist models trained separately on the corresponding datasets

Dataset	Anatomical Targets	DSC↑		NSD↑	
		nnUNet	SAT-Nano	nnUNet	SAT-Nano
AbdomenCT1K [49]	Kidney, liver, pancreas, spleen	95.09	94.22	88.24	87.78
ACDC [7]	Left and right ventricles, myocardium	90.76	84.22	97.62	66.44
Brain Atlas [57]	Anatomical brain atlas	83.78	72.25	85.35	68.57
BrainPTM [4]	Brain anatomies	68.37	60.04	33.90	46.93
BraTS2021 [5]	Brain tumor	73.37	68.54	68.66	66.14
Challenge 4C2021	Laryngeal or hypopharyngeal cancer	52.96	30.87	50.02	30.17
CHAOS CT [33]	Liver	97.08	96.83	81.04	83.50
CHAOS MR [33]	Kidney, liver, spleen	88.80	82.33	64.35	50.04
Couinaud [59]	Liver Couinaud segments	87.86	79.67	70.44	53.33
COVID-19 CT Seg [47]	Lung and COVID-19 infection	91.53	70.21	77.02	49.47
FLARE22 [48]	Abdominal organs	93.36	89.13	91.15	87.43
HAN Seg [54]	Head and neck anatomies	62.18	71.04	63.52	76.68
ISLES2022 [29]	Brain stroke	64.95	42.37	60.10	40.97
KITS21 [28]	Kidney, kidney tumor, kidney cyst	77.90	67.12	73.30	60.27
LUNA16 [58]	Lung and trachea	96.64	96.34	93.85	94.50
MRSpineSeg [52]	Thoracic and lumbar vertebrae, sacrum	68.97	67.99	58.82	60.93
MSD Cardiac [3]	Left atrium	94.28	89.38	64.21	73.33
MSD Colon [3]	Colon cancer	54.39	36.53	51.07	30.98
MSD HepaticVessel [3]	Hepatic vessels and liver tumor	67.74	59.93	61.85	53.80
MSD Hippocampus [3]	Hippocampus	89.18	87.15	97.92	95.88
MSD Liver [3]	Liver and liver tumor	77.92	79.66	63.78	66.86
MSD Lung [3]	Lung cancer	71.74	57.65	58.32	47.37
MSD Pancreas [3]	Pancreas and pancreas tumor	68.64	55.83	53.31	46.53
MSD Prostate [3]	Prostate	71.32	69.79	50.19	50.48
MSD Spleen [3]	Spleen	92.95	94.11	88.01	83.90
NSCLC [6]	Thoracic cavity and effusion	42.20	76.37	38.63	61.87
PROMISE12 [42]	Prostate	88.86	82.31	71.90	54.60
SKI10 [38]	Femur and tibia	88.15	81.38	94.38	88.23
SLIVER07 [26]	Liver	97.30	97.33	97.72	87.62
TotalSegmentator [64]	Whole body organs and tissues	93.30	92.66	84.83	92.07
WMH [36]	White matter hyper-intensities	77.02	64.24	88.88	78.10
WORD [45]	Abdominal organs	85.49	88.12	78.79	79.84

Table 6 | Segmentation results on lesions. BT: brain tumor, CC: colon cancer, C-19: COVID-19 infection, PE: pleural effusion, KT: kidney tumor, L/HC: Laryngeal/Hypopharyngeal Cancer, LT: liver tumor, LC: Lung Cancer, PT: pancreas tumor.

Metric	Method	BT	CC	C-19	PE	KT	L/HC	LT	lC	PT	Stroke
DSC↑	nnUNet	88.01	54.39	80.97	13.83	86.64	52.96	68.38	71.74	50.83	64.95
	SAT-Nano	86.00	36.53	35.30	57.14	71.16	30.87	62.56	57.65	23.90	42.37
NSD↑	nnUNet	74.09	51.07	57.38	24.80	73.84	50.02	50.00	58.32	38.14	60.10
	SAT-Nano	73.13	30.98	18.12	49.03	55.12	30.17	46.24	47.37	16.64	40.97

References

[1] Nikolov, stanislav and blackwell, sam and zverovitch, alexei and mendes, ruheena and livne, michelle and de fauw, jeffrey and patel, yojan and meyer, clemens and askham, harry and romera-paredes, bernadino and kelly, christopher and karthikesalingam, alan and chu, carlton and carnell, dawn and boon, cheng and d'souza, derek and moinuddin, syed ali and garie, bethany and mcquinlan, yasmin and ireland, sarah and hampton, kiarna and fuller, krystle and montgomery, hugh and rees, geraint and suleyman, mustafa and back, trevor and hughes, cían owen and ledsam, joseph r and ronneberger, olaf. *Journal of Medical Internet Research*, 23(7):e26151, 2021.

[2] OpenAI (2023). Gpt-4 technical report, 2023.

[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] Itzik Avital, Ilya Nelkenbaum, Galia Tsarfaty, Eli Konen, Nahum Kiryati, and Arnaldo Mayer. Neural segmentation of seeding rois (srois) for pre-surgical brain tractography. *IEEE Transactions on Medical Imaging*, 39(5):1655–1667, 2019.

[5] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

[6] Shaimaa Bakr, Olivier Gevaert, Sebastian Echegaray, Kelsey Ayers, Mu Zhou, Majid Shafiq, Hong Zheng, Jalen Anthony Benson, Weiruo Zhang, Ann NC Leung, et al. A radiogenomic dataset of non-small cell lung cancer. *Scientific Data*, 5(1):1–9, 2018.

[7] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018.

[8] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

[9] Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Universeg: Universal medical image segmentation. *arXiv preprint arXiv:2304.06131*, 2023.

[10] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision*, pages 205–218. Springer, 2022.

[11] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng Ann Heng. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *IEEE Transactions on Medical Imaging*, 39(7):2494–2505, 2020.

[12] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, et al. 3d transunet: Advancing medical image segmentation through vision transformers. *arXiv preprint arXiv:2310.07781*, 2023.

[13] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023.

[14] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 424–432. Springer, 2016.

[15] Hui Cui, Yiyue Xu, Wanlong Li, Linlin Wang, and Henry Duh. Collaborative learning of cross-channel clinical attention for radiotherapy-related esophageal fistula prediction from ct. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 212–220. Springer, 2020.

[16] Jiahua Dong, Yang Cong, Gan Sun, Bineng Zhong, and Xiaowei Xu. What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4023–4032, 2020.

[17] Qi Dou, Lequan Yu, Hao Chen, Yueming Jin, Xin Yang, Jing Qin, and Pheng-Ann Heng. 3d deeply supervised network for automated segmentation of volumetric medical images. *Medical Image Analysis*, 41:40–54, 2017.

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

[19] Leyuan Fang, Chong Wang, Shutao Li, Hossein Rabbani, Xiangdong Chen, and Zhimin Liu. Attention to lesion: Lesion-aware convolutional neural network for retinal optical coherence tomography image classification. *IEEE Transactions on Medical Imaging*, 38(8):1959–1970, 2019.

[20] Xi Fang and Pingkun Yan. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. *IEEE Transactions on Medical Imaging*, 39(11):3619–3629, 2020.

[21] Eli Gibson, Francesco Giganti, Yipeng Hu, Ester Bonmati, Steve Bandula, Kurinchi Gurusamy, Brian Davidson, Stephen P Pereira, Matthew J Clarkson, and Dean C Barratt. Automatic multi-organ segmentation on abdominal ct with dense v-networks. *IEEE Transactions on Medical Imaging*, 37(8):1822–1834, 2018.

[22] 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.

[23] Ivan Gonzalez-Diaz. Dermaknet: Incorporating the knowledge of dermatologists to convolutional neural networks for skin lesion diagnosis. *IEEE journal of biomedical and health informatics*, 23(2):547–559, 2018.

[24] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2021.

[25] 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.

[26] Tobias Heimann, Bram Van Ginneken, Martin A Styner, Yulia Arzhaeva, Volker Aurich, Christian Bauer, Andreas Beck, Christoph Becker, Reinhard Beichel, György Bekes, et al. Comparison and evaluation of methods for liver segmentation from ct datasets. *IEEE Transactions on Medical Imaging*, 28(8):1251–1265, 2009.

[27] Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, page 101821, 2020.

[28] Nicholas Heller, Fabian Isensee, Dasha Trofimova, Resha Tejpaul, Zhongchen Zhao, Huai Chen, Lisheng Wang, Alex Golts, Daniel Khapun, Daniel Shats, et al. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct. *arXiv preprint arXiv:2307.01984*, 2023.

[29] Moritz R Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, et al. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific Data*, 9(1):762, 2022.

[30] 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? *arXiv preprint arXiv:2304.14660*, 2023.

[31] 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.

[32] Zhanghexuan Ji, Dazhou Guo, Puyang Wang, Ke Yan, Le Lu, Minfeng Xu, Qifeng Wang, Jia Ge, Mingchen Gao, Xianghua 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.

[33] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021.

[34] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NACCL-HLT)*, pages 4171–4186, 2019.

[35] 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. *arXiv preprint arXiv:2304.02643*, 2023.

[36] Hugo J Kuijf, J Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adria Casamitjana, et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE Transactions on Medical Imaging*, 38(11):2556–2568, 2019.

[37] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[38] Soochahn Lee, Hackjoon Shim, Sang Hyun Park, Il Dong Yun, and Sang Uk Lee. Learning local shape and appearance for segmentation of knee cartilage in 3d mri. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 231–240, 2010.

[39] Jiayu Lei, Lisong Dai, Haoyun Jiang, Chaoyi Wu, Xiaoman Zhang, Yao Zhang, Jiangchao Yao, Weidi Xie, Yanyong Zhang, Yuehua Li, Ya Zhang, and Yanfeng Wang. Unibrain: Universal brain mri diagnosis with hierarchical knowledge-enhanced pre-training. *arXiv preprint arXiv:2309.06828*, 2023.

[40] Liu Li, Mai Xu, Xiaofei Wang, Lai Jiang, and Hanruo Liu. Attention based glaucoma detection: A large-scale database and cnn model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10571–10580, 2019.

[41] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE Transactions on Medical Imaging*, 37(12):2663–2674, 2018.

[42] Geert Litjens, Robert Toth, Wendy Van De Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram Van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical Image Analysis*, 18(2):359–373, 2014.

[43] 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.

[44] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[45] X Luo, W Liao, J Xiao, J Chen, T Song, X Zhang, K Li, DN Metaxas, G Wang, and S Zhang. Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *Medical Image Analysis*, 82:102642–102642, 2022.

[46] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023.

[47] Jun Ma, Yixin Wang, Xingle An, Cheng Ge, Ziqi Yu, Jianan Chen, Qiongjie Zhu, Guoqiang Dong, Jian He, Zhiqiang He, et al. Toward data-efficient learning: A benchmark for covid-19 ct lung and infection segmentation. *Medical physics*, 48(3):1197–1210, 2021.

[48] Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Ma, Adamo Young, Cheng Zhu, Kangkang Meng, Xin Yang, Ziyan Huang, Fan Zhang, Wentao Liu, YuanKe Pan, Shoujin Huang, Jiacheng Wang, Mingze Sun, Weixin Xu, Dengqiang Jia, Jae Won Choi, Natália Alves, Bram de Wilde, Gregor Koehler, Yajun Wu, Manuel Wiesenfarth, Qiongjie Zhu, Guoqiang Dong, Jian He, the FLARE Challenge Consortium, and Bo Wang. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. *arXiv preprint arXiv:2308.05862*, 2023.

[49] 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.

[50] Lena Maier-Hein, Bjoern Menze, et al. Metrics reloaded: Pitfalls and recommendations for image analysis validation. *arXiv. org*, (2206.01653), 2022.

[51] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision (3DV)*, pages 565–571. Ieee, 2016.

[52] Shumao Pang, Chunlan Pang, Lei Zhao, Yangfan Chen, Zhihai Su, Yujia Zhou, Meiyan Huang, Wei Yang, Hai Lu, and Qianjin Feng. Spineparsenet: spine parsing for volumetric mr image by a two-stage segmentation framework with semantic image representation. *IEEE Transactions on Medical Imaging*, 40(1):262–273, 2020.

[53] Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A Silva. Brain tumor segmentation using convolutional neural networks in mri images. *IEEE Transactions on Medical Imaging*, 35(5):1240–1251, 2016.

[54] Gašper Podobnik, Primož Strojan, Primož Peterlin, Bulat Ibragimov, and Tomaž Vrtovec. Han-seg: The head and neck organ-at-risk ct and mr segmentation dataset. *Medical physics*, 50(3):1917–1927, 2023.

[55] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.

[56] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*, 53:197–207, 2019.

[57] Ahmed Serag, Paul Aljabar, Gareth Ball, Serena J Counsell, James P Boardman, Mary A Rutherford, A David Edwards, Joseph V Hajnal, and Daniel Rueckert. Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression. *Neuroimage*, 59(3):2255–2265, 2012.

[58] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42:1–13, 2017.

[59] Jiang Tian, Li Liu, Zhongchao Shi, and Feiyu Xu. Automatic couinaud segmentation from ct volumes on liver using glc-unet. In *International Workshop on Machine Learning in Medical Imaging*, pages 274–282. Springer, 2019.

[60] Constantin Ulrich, Fabian Isensee, Tassilo Wald, Maximilian Zenk, Michael Baumgartner, and Klaus H Maier-Hein. Multitalent: A multi-dataset approach to medical image segmentation. *arXiv preprint arXiv:2303.14444*, 2023.

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[62] Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyan Huang, Yiqing Shen, et al. Sam-med3d. *arXiv preprint arXiv:2310.15161*, 2023.

[63] Yixin Wang, Yao Zhang, Yang Liu, Jiang Tian, Cheng Zhong, Zhongchao Shi, Yang Zhang, and Zhiqiang He. Does non-covid-19 lung lesion help? investigating transferability in covid-19 ct image segmentation. *Computer Methods and Programs in Biomedicine*, 202:106004, 2021.

[64] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5), 2023.

[65] Chaoyi Wu, Feng Chang, Xiao Su, Zhihan Wu, Yanfeng Wang, Ling Zhu, and Ya Zhang. Integrating features from lymph node stations for metastatic lymph node detection. *Computerized Medical Imaging and Graphics*, 101:102108, 2022.

[66] Chaoyi Wu, Xiaoman Zhang, Yanfeng Wang, Ya Zhang, and Weidi Xie. K-diag: Knowledge-enhanced disease diagnosis in radiographic imaging. *MICCAI Workshop on Big Task Small Data (BTSD)*, 2023.

[67] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21372–21383, October 2023.

[68] Xiaozheng Xie, Jianwei Niu, Xuefeng Liu, Zhengsu Chen, Shaojie Tang, and Shui Yu. A survey on incorporating domain knowledge into deep learning for medical image analysis. *Medical Image Analysis*, 69:101985, 2021.

[69] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 171–180. Springer, 2021.

[70] Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, and Yong Xia. Uniseg: A prompt-driven universal segmentation model as well as a strong representation learner. *arXiv preprint arXiv:2304.03493*, 2023.

[71] Qihang Yu, Lingxi Xie, Yan Wang, Yuyin Zhou, Elliot K Fishman, and Alan L Yuille. Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8280–8289, 2018.

[72] Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1):4542, 2023.

[73] Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 107–117. Springer, 2022.

[74] Yixiao Zhang, Xinyi Li, Huimiao Chen, Alan L Yuille, Yaoyao Liu, and Zongwei Zhou. Continual learning for abdominal multi-organ and tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 35–45. Springer, 2023.

[75] Yao Zhang, Jiawei Yang, Yang Liu, Jiang Tian, Siyun Wang, Cheng Zhong, Zhongchao Shi, Yang Zhang, and Zhiqiang He. Decoupled pyramid correlation network for liver tumor segmentation from ct images. *Medical Physics*, 49(11):7207–7221, 2022.

[76] Yao Zhang, Jiawei Yang, Jiang Tian, Zhongchao Shi, Cheng Zhong, Yang Zhang, and Zhiqiang He. Modality-aware mutual learning for multi-modal medical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 589–599. Springer, 2021.

[77] Qiaoyu Zheng, Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Large-scale long-tailed disease diagnosis on radiology images. *arXiv preprint arXiv:2312.16151*, 2023.

[78] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: volumetric medical image segmentation via a 3d transformer. *IEEE Transactions on Image Processing*, 2023.

[79] Yuhang Zhou, Shu-Wen Sun, Qiu-Ping Liu, Xun Xu, Ya Zhang, and Yu-Dong Zhang. Ted: Two-stage expert-guided interpretable diagnosis framework for microvascular invasion in hepatocellular carcinoma. *Medical Image Analysis*, 82:102575, 2022.

[80] Yuhang Zhou, Xiaoman Zhang, Shixiang Feng, and Ya Zhang. Uncertainty-aware incremental learning for multi-organ segmentation. *arXiv preprint arXiv:2103.05227*, 2021.

[81] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6):1856–1867, 2019.

9 Appendix

9.1 Details of SAT-DS

Fig. 7 shows the distribution of classes and volumes in the SAT-DS.

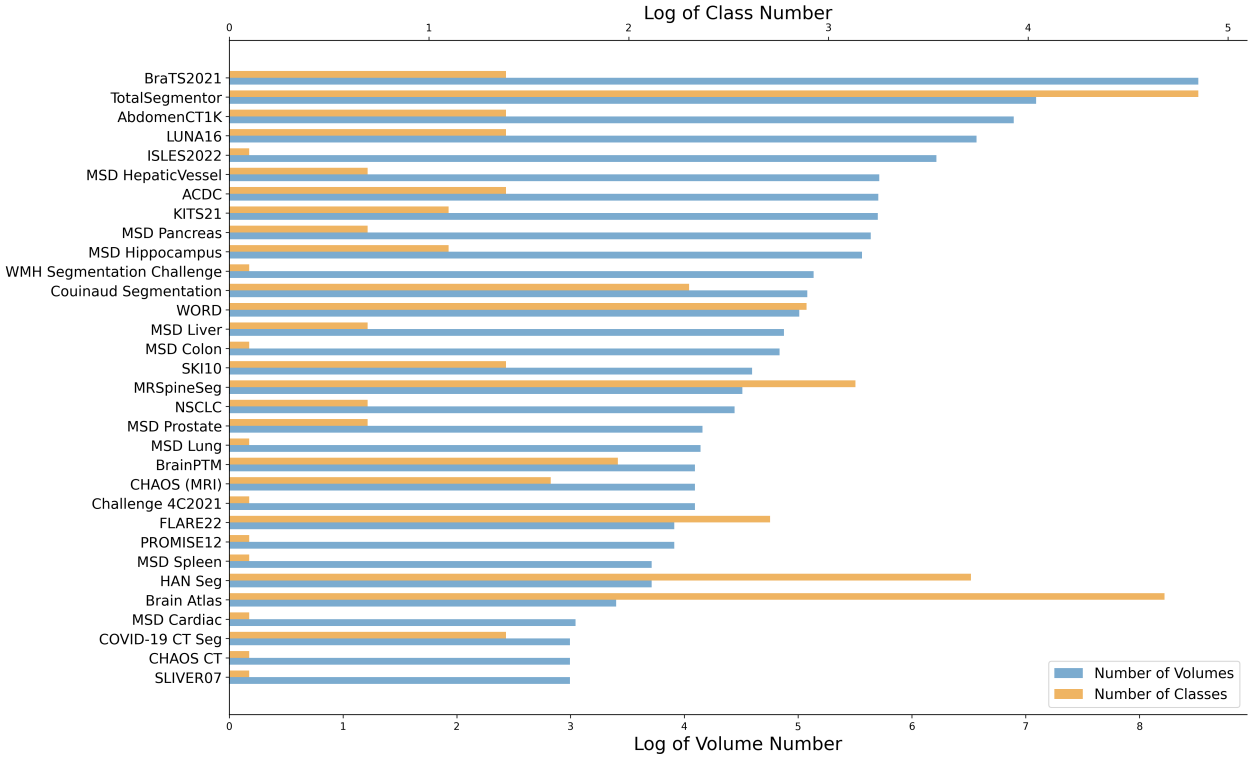

Figure 7 | Example statistics of SAT-DS dataset. The number of volumes and annotations are marked with blue and orange bar respectively. To facilitate presentation, logarithms are applied to all number.

9.2 Detailed architecture of nnU-Nets specialist

Table 7 presents the architecture of the nnU-Nets specialist trained on each datasets.

Table 7 | The configuration of each nnU-Net trained on the corresponding dataset.

Dataset	#Class	Train/Test	Input Size	#Stage	#Depth	#Width	Model Size
AbdomenCT1K [49]	4	790/198	[96 160 160]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
ACDC [7]	3	200/100	[10 256 224]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	30M
BrainAtlas [57]	95	24/6	[112 128 112]	5	[2 2 2 2 2]	[32 64 128 256 320]	17M
BrainPTM [4]	5	48/12	[112 144 112]	5	[2 2 2 2 2]	[32 64 128 256 320]	17M
BraTS2021 [5]	3	4004/1000	[128 128 128]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
Challenge4C2021	1	48/12	[40 224 192]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
CHAOSCT [33]	1	16/4	[48 224 192]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
CHAOSMR [33]	4	48/12	[32 192 288]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
Couinaud [59]	8	126/35	[64 192 192]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
COVID19 [47]	3	16/4	[56 192 192]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
FLARE22 [48]	13	40/10	[40 224 192]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
HANSeg [54]	30	33/8	[40 224 192]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
ISLES2022 [29]	1	400/100	[80 96 80]	5	[2 2 2 2 2]	[32 64 128 256 320]	17M
KiTS21 [28]	3	239/60	[128 128 128]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
LUNA16 [58]	3	710/178	[80 192 160]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
MRSpineSeg [52]	19	73/18	[8 640 320]	7	[2 2 2 2 2 2 2]	[32 64 128 256 320 320 320]	43M
MSD Heart [3]	1	16/4	[80 192 160]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
MSD Colon [3]	1	101/25	[56 192 192]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
MSD HepaticVessel [3]	2	242/61	[64 192 192]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
MSD Hippocampus [3]	2	208/52	[40 56 40]	4	[2 2 2 2]	[32 64 128 256]	6M
MSD Liver [3]	2	105/26	[128 128 128]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
MSD Lung [3]	1	50/13	[80 192 160]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
MSD Pancreas [3]	2	225/56	[40 224 224]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
MSD Prostate [3]	2	52/12	[16 320 320]	7	[2 2 2 2 2 2 2]	[32 64 128 256 320 320 320]	45M
MSD Spleen [3]	1	33/8	[64 192 160]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
NSCLC [6]	2	68/17	[48 224 192]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
PROMISE12 [42]	1	40/10	[20 320 256]	7	[2 2 2 2 2 2 2]	[32 64 128 256 320 320 320]	45M
SKI10 [38]	4	59/40	[64 192 160]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
SLIVER07 [26]	1	16/4	[80 192 160]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
TotalSegmentator Heart [64]	13	962/240	[128 128 128]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
TotalSegmentator Muscles [64]	21	962/240	[128 128 128]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
TotalSegmentator Organs [64]	20	962/240	[128 128 128]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
TotalSegmentator Ribs [64]	24	962/240	[128 128 128]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
TotalSegmentator Vertebrae [64]	25	962/240	[128 128 128]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
WMH [36]	1	60/110	[48 224 192]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M
WORD [45]	16	100/50	[64 192 160]	6	[2 2 2 2 2 2]	[32 64 128 256 320 320]	31M

9.3 All classes in SAT-DS

Table 8 illustrates the 362 anatomies from CT or MR images in our SAT-DS.

Table 8 | Class Name List. We show the detailed class name list in our SAT-DS.

Name List 1	Name List 2	Name List 3
ct_liver	ct_right scapula	mri_right subgenual frontal cortex
ct_kidney	ct_autochthon	mri_right cuneus
ct_spleen	ct_clavicula	mri_left posterior temporal lobe
ct_pancreas	ct_femur	mri_left precentral gyrus
ct_right kidney	ct_gluteus maximus	mri_left lateral orbital gyrus
ct_aorta	ct_gluteus medius	mri_right cerebellum
ct_inferior vena cava	ct_gluteus minimus	mri_right putamen
ct_right adrenal gland	ct_hip	mri_right lateral orbital gyrus
ct_left adrenal gland	ct_humerus	mri_left lingual gyrus
ct_gallbladder	ct_iliopsoas	mri_right postcentral gyrus
ct_esophagus	ct_scapula	mri_left lateral remainder occipital lobe
ct_stomach	ct_colon cancer	mri_corpus callosum
ct_duodenum	mri_left heart atrium	mri_left subgenual frontal cortex
ct_left kidney	ct_kidney cyst	mri_right superior frontal gyrus
ct_adrenal gland	mri_brain tumor	mri_right superior temporal gyrus middle part
ct_kidney tumor	mri_transition zone of prostate	mri_right lateral remainder occipital lobe
ct_colon	mri_peripheral zone of prostate	mri_left amygdala
ct_intestine	ct_thoracic cavity	mri_right lateral ventricle temporal horn
ct_rectum	ct_lung effusion	mri_right middle and inferior temporal gyrus
ct_bladder	mri_liver	mri_right medial orbital gyrus
ct_lung tumor	mri_right kidney	mri_left thalamus
ct_left lung lower lobe	mri_left kidney	mri_left superior temporal gyrus anterior part
ct_right lung lower lobe	mri_spleen	mri_right anterior orbital gyrus
ct_right lung middle lobe	mri_kidney	mri_third ventricle
ct_left lung upper lobe	ct_arytenoid	mri_left cerebellum
ct_right lung upper lobe	ct_brainstem	mri_right subcallosal area
ct_small bowel	ct_buccal mucosa	mri_right fusiform gyrus
ct_trachea	ct_left carotid artery	mri_right middle frontal gyrus
ct_urinary bladder	ct_right carotid artery	mri_right insula middle short gyrus
ct_lung lower lobe	ct_cervical esophagus	mri_right insula anterior inferior cortex
ct_lung upper lobe	ct_left cochlea	mri_left anterior temporal lobe lateral part
ct_sacrum	ct_right cochlea	mri_right insula posterior short gyrus
ct_cervical vertebrae 1 (c1)	ct_cricopharyngeal inlet	mri_left anterior temporal lobe medial part
ct_cervical vertebrae 2 (c2)	ct_left anterior eyeball	mri_right lateral ventricle excluding temporal horn
ct_cervical vertebrae 3 (c3)	ct_right anterior eyeball	mri_right precentral gyrus
ct_cervical vertebrae 4 (c4)	ct_left posterior eyeball	mri_left angular gyrus
ct_cervical vertebrae 5 (c5)	ct_right posterior eyeball	mri_right pre-subgenual frontal cortex
ct_cervical vertebrae 6 (c6)	ct_left lacrimal gland	mri_hippocampus
ct_cervical vertebrae 7 (c7)	ct_right lacrimal gland	mri_brainstem excluding substantia nigra
ct_lumbar vertebrae 1 (l1)	ct_larynx - glottis	mri_left substantia nigra
ct_lumbar vertebrae 2 (l2)	ct_larynx - supraglottic	mri_right thalamus
ct_lumbar vertebrae 3 (l3)	ct_lips	mri_right hippocampus
ct_lumbar vertebrae 4 (l4)	ct_mandible	mri_left pre-subgenual frontal cortex
ct_lumbar vertebrae 5 (l5)	ct_optic chiasm	mri_left anterior orbital gyrus

Table 8 | Class Name List. We show the detailed class name list in our SAT-DS

Name List 1	Name List 2	Name List 3
ct_thoracic vertebrae 1 (t1)	ct_left optic nerve	mri_right anterior temporal lobe medial part
ct_thoracic vertebrae 10 (t10)	ct_right optic nerve	mri_left insula anterior inferior cortex
ct_thoracic vertebrae 11 (t11)	ct_oral cavity	mri_left inferior frontal gyrus
ct_thoracic vertebrae 12 (t12)	ct_left parotid gland	mri_right parahippocampal and ambient gyrus
ct_thoracic vertebrae 2 (t2)	ct_right parotid gland	mri_left medial orbital gyrus
ct_thoracic vertebrae 3 (t3)	ct_pituitary gland	mri_right nucleus accumbens
ct_thoracic vertebrae 4 (t4)	ct_spinal cord	mri_right lingual gyrus
ct_thoracic vertebrae 5 (t5)	ct_left submandibular gland	mri_right superior temporal gyrus anterior part
ct_thoracic vertebrae 6 (t6)	ct_right submandibular gland	mri_right superior parietal gyrus
ct_thoracic vertebrae 7 (t7)	ct_thyroid	mri_left posterior orbital gyrus
ct_thoracic vertebrae 8 (t8)	ct_carotid artery	mri_left supramarginal gyrus
ct_thoracic vertebrae 9 (t9)	ct_cochlea	mri_left insula anterior short gyrus
ct_cervical vertebrae	ct_anterior eyeball	mri_cerebellum
ct_lumbar vertebrae	ct_posterior eyeball	mri_right substantia nigra
ct_thoracic vertebrae	ct_lacrimal gland	mri_right posterior temporal lobe
ct_left heart atrium	ct_optic nerve	mri_left putamen
ct_right heart atrium	ct_parotid gland	mri_left cuneus
ct_heart myocardium	ct_submandibular gland	mri_right pallidum
ct_left heart ventricle	ct_eyeball	mri_left subcallosal area
ct_right heart ventricle	mri_stroke	mri_amygdala
ct_left iliac artery	ct_left lung	mri_right posterior cingulate gyrus
ct_right iliac artery	ct_right lung	mri_right anterior cingulate gyrus
ct_left iliac vena	ct_covid-19 infection	mri_left postcentral gyrus
ct_right iliac vena	mri_left optic radiation	mri_left superior parietal gyrus
ct_portal vein and splenic vein	mri_right optic radiation	mri_left insula middle short gyrus
ct_pulmonary artery	mri_left corticospinal tract	mri_left fusiform gyrus
ct_heart atrium	mri_right corticospinal tract	mri_left insula anterior long gyrus
ct_heart ventricle	mri_optic radiation	mri_left hippocampus
ct_iliac artery	mri_corticospinal tract	mri_left insula posterior short gyrus
ct_iliac vena	mri_left ventricle cavity	mri_basal ganglia
ct_left rib 1	mri_right ventricle cavity	mri_insula
ct_left rib 10	mri_myocardium	mri_left superior temporal gyrus middle part
ct_left rib 11	mri_ventricle cavity	mri_right amygdala
ct_left rib 12	ct_laryngeal cancer or hypopharyngeal cancer	mri_right angular gyrus
ct_left rib 2	mri_anterior hippocampus	mri_parietal lobe
ct_left rib 3	mri_posterior hippocampus	mri_brainstem
ct_left rib 4	mri_sacrum	mri_right insula anterior long gyrus
ct_left rib 5	mri_lumbar vertebrae 5 (l5)	mri_frontal lobe
ct_left rib 6	mri_lumbar vertebrae 4 (l4)	mri_occipital lobe
ct_left rib 7	mri_lumbar vertebrae 3 (l3)	mri_right caudate nucleus
ct_left rib 8	mri_lumbar vertebrae 2 (l2)	mri_lateral ventricle
ct_left rib 9	mri_lumbar vertebrae 1 (l1)	mri_thalamus
ct_right rib 1	mri_thoracic vertebrae 12 (t12)	mri_left pallidum
ct_right rib 10	mri_thoracic vertebrae 11 (t11)	mri_left nucleus accumbens
ct_right rib 11	mri_thoracic vertebrae 10 (t10)	mri_right insula anterior short gyrus
ct_right rib 12	mri_thoracic vertebrae 9 (t9)	mri_right supramarginal gyrus
ct_right rib 2	mri_intervertebral disc between lumbar vertebrae 5 (l5) and sacrum	mri_left middle and inferior temporal gyrus
ct_right rib 3	mri_intervertebral disc between lumbar vertebrae 4 (l4) and lumbar vertebrae 5 (l5)	mri_right inferior frontal gyrus

Table 8 | Class Name List. We show the detailed class name list in our SAT-DS

Name List 1	Name List 2	Name List 3
ct_right rib 4	mri_intervertebral disc between lumbar vertebrae 3 (l3) and lumbar vertebrae 4 (l4)	mri_right anterior temporal lobe lateral part
ct_right rib 5	mri_intervertebral disc between lumbar vertebrae 2 (l2) and lumbar vertebrae 3 (l3)	mri_temporal lobe
ct_right rib 6	mri_intervertebral disc between lumbar vertebrae 1 (l1) and lumbar vertebrae 2 (l2)	mri_right insula posterior long gyrus
ct_right rib 7	mri_intervertebral disc between thoracic vertebrae 12 (t12) and lumbar vertebrae 1 (l1)	mri_left superior frontal gyrus
ct_right rib 8	mri_intervertebral disc between thoracic vertebrae 11 (t11) and thoracic vertebrae 12 (t12)	mri_left posterior cingulate gyrus
ct_right rib 9	mri_intervertebral disc between thoracic vertebrae 10 (t10) and thoracic vertebrae 11 (t11)	mri_left middle frontal gyrus
ct_right rib	mri_intervertebral disc between thoracic vertebrae 9 (t9) and thoracic vertebrae 10 (t10)	mri_left lateral ventricle excluding temporal horn
ct_left rib	mri_lumbar vertebrae	mri_cingulate gyrus
ct_rib	mri_thoracic vertebrae	mri_left caudate nucleus
ct_left autochthon	mri_intervertebral discs	ct_vertebrae
ct_right autochthon	mri_prostate	mri_vertebrae
ct_brain	ct_lung	ct_pancreas tumor
ct_left clavicula	mri_femur bone	ct_liver tumor
ct_right clavicula	mri_femur cartilage	mri_non-enhancing brain tumor
ct_left femur	mri_tibia bone	mri_enhancing brain tumor
ct_right femur	mri_tibia cartilage	mri_brain
ct_left gluteus maximus	mri_white matter hyperintensities	ct_left eyeball
ct_right gluteus maximus	ct_head of left femur	ct_right eyeball
ct_left gluteus medius	ct_head of right femur	ct_caudate lobe
ct_right gluteus medius	ct_head of femur	ct_left lateral superior segment of liver
ct_left gluteus minimus	ct_liver vessel	ct_left lateral inferior segment of liver
ct_right gluteus minimus	mri_brain edema	ct_left medial segment of liver
ct_left hip	mri_right posterior orbital gyrus	ct_right anterior inferior segment of liver
ct_right hip	mri_left parahippocampal and ambient gyrus	ct_right posterior inferior segment of liver
ct_left humerus	mri_left insula posterior long gyrus	ct_right posterior superior segment of liver
ct_right humerus	mri_left lateral ventricle temporal horn	ct_right anterior superior segment of liver
ct_left iliopsoas	mri_left anterior cingulate gyrus	ct_left lobe of liver
ct_right iliopsoas	mri_left straight gyrus	ct_right lobe of liver
ct_left scapula	mri_right straight gyrus	