




Supplementary of SAM-Med3D

Haoyu Wang^{1,2}, Sizheng Guo¹, Jin Ye¹, Zhongying Deng¹, Junlong Cheng¹,
Tianbin Li¹, Jianpin Chen¹, Yanzhou Su¹, Ziyang Huang^{1,2}, Yiqing Shen¹, Bin
Fu³, Shaoting Zhang¹, Junjun He¹, and Yu Qiao¹^{*}

¹ Shanghai Jiao Tong University

² Shanghai Artificial Intelligence Laboratory

³ Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
hejunjun@pjlab.org.cn

Table of contents

- §A: Additional Dataset Details
- §B: SAM-Med3D Model Details
- §C: Additional Experiment Details

A Additional Dataset Details

A.1 Training Datasets

As introduced in the main body, we adopt a 2-stage training procedure for SAM-Med3D. In the 1st-stage pre-training, we use the full training set of SAM-Med3D-140K (58 public datasets and all private datasets); in the 2nd-stage fine-tuning, we use a selected version of the training set (44 datasets). The detailed list of which public datasets are used in our training set and whether they are chosen for fine-tuning is in Tab. 1.

A.2 Evaluation Datasets

We selected 16 datasets for our evaluations. The 16 datasets are categorized into three groups to facilitate a more focused analysis.

Seen organ segmentation benchmark includes *Totalseg-Test* (official test set of *Totalsegmentator* [16]), *AMOS-Val* (official validation set of *AMOS* [5]), *BTCV* [6]. In addition to the datasets with official splits, we utilized all the labeled data from *BTCV* for testing. These three datasets jointly offer masks for more than 100 kinds of organs across different anatomical structures throughout the body, covering both CT and MRI modalities.

Unseen organ segmentation benchmark contains *HaN-Seg* [9], *FeTA21* [8], *FeTA22* [8], *iSeg17* [15], *iSeg19* [12], *MRBrains13* [7], *MRBrains18* [2], *cSeg22* [13], and *SEG.A.2023* [11]. These nine datasets collectively provide annotations of

^{*} Corresponding authors.

Table 1: Details of the training set of SA-Med3D-140K.

dataset	pre-train	fine-tune	dataset	pre-train	fine-tune
AbdomenCT-1K	✓	✓	KiTS21	✓	✓
AMOS22	✓	✓	LAScarQS22	✓	✓
ASC18	✓		LITS	✓	✓
ATM22	✓	✓	LUNA16	✓	
BrainPTM	✓	✓	MMWHS	✓	✓
BraTS13	✓		MSD_BrainTumour	✓	✓
BraTS15	✓		MSD_Colon	✓	✓
BraTS18	✓		MSD_Heart	✓	
BraTS19	✓		MSD_HepaticVessel	✓	✓
BraTS20	✓	✓	MSD_Hippocampus	✓	
BraTS21	✓	✓	MSD_Liver	✓	✓
BTCV_Cervix	✓	✓	MSD_Lung	✓	✓
CAUSE07	✓	✓	MSD_Pancreas	✓	✓
CHAOS	✓	✓	MSD_Prostate	✓	
COSMOS22	✓	✓	MSD_Spleen	✓	✓
COVID19CTscans	✓	✓	Parse22	✓	✓
CrossMoDA21	✓		PICAI	✓	
CrossMoDA22	✓		Promise09	✓	✓
CT_ORG	✓	✓	Promise12	✓	✓
CTPelvic1k	✓	✓	ProstateMRISeg	✓	✓
FLARE21	✓	✓	RibFrac2020	✓	
FLARE22	✓	✓	SegThor	✓	✓
HeartSegMRI	✓	✓	SLIVER07	✓	✓
hvsmr16	✓		STACOM_SLAWT	✓	✓
ISLESSISS	✓	✓	StructSeg19	✓	
ISLESSPES	✓	✓	Totalsegmentator	✓	✓
VerSe19	✓	✓	KIPA22	✓	✓
VerSe20	✓	✓	KiTS19	✓	✓
VESSEL12	✓	✓	WORD	✓	✓
24 Private datasets	✓				

various complex unseen organs, including the gland, cerebellum and white matter. All the labeled data from these datasets are utilized for testing and the categories are unseen in training. Notably, *HaN-Seg* and *SEG.A.2023* provides masks in CT images and other seven datasets provide MRI images. These datasets together constitute a challenging benchmark due to their varied data sources. Besides, some categories like brain and brainstem are seen in training.

Lesion segmentation benchmark consists of *KiTS21-Val* (official validation set of *KiTS21* [4]), *BraTS21-Val** (our validation set of *BraTS21* [1]), *ATLAS* [10] and *TDSC-ABUS* [14]. For brain tumor evaluation, we randomly selected 90 cases from *BraTS21* [1] to construct *BraTS21-Val**, maintaining consistent data scale with *KiTS21-Val*. Regarding the additional dataset from MIC-CAI2023 Challenge (i.e. *ATLAS* and *TDSC-ABUS*), all of their labeled data are

used for evaluation. It is paramount to note that these two datasets offer annotations for tumor types not encountered during training (i.e. unseen targets), with *TDSC-ABUS* introducing an unseen modality: US (Ultrasound). Both *ATLAS* and *TDSC-ABUS* are instrumental in gauging the method’s generalization capabilities.

A.3 Anatomical Structure Taxonomy

To better present our experimental results, we reorganized them using a taxonomy based on anatomical structures. To clarify our taxonomic criteria and facilitate understanding, we have enumerated the target categories corresponding to each anatomical structure in the Table 2. Additionally, the ‘abdomen&thorax’ category in our taxonomy does not strictly adhere to these two regions alone; it also encompasses organs from the Head and Neck area. Conversely, organs that are classified under cardiac, gland, or muscle, even if located in the thoracoabdominal area, are not categorized under abdomen&thorax in our taxonomy.

Table 2: Anatomical structures and their major targets.

Anatomical Structure	Major Targets
abdomen&thorax	aorta, bladder, carotid artery, colon, duodenum, esophagus, eyeball, gallbladder, inferior vena cava, kidney, liver, lung, pancreas, portal vein and splenic vein, prostate and uterus, pulmonary artery, small bowel, spleen, stomach, trachea, urinary bladder, adrenal gland
bone	clavícula, femur, hip, humerus, rib (L1-12, R1-12), sacrum, scapula, vertebrae (C1-7, L1-5, T1-12)
brain	brain, brainstem
cardiac	heart atrium, heart myocardium, heart ventricle
muscle	autochthon, gluteus muscle, iliopsoas
lesion	edema, enhancing tumor, kidney tumor
unseen organ	parotid gland, submandibular gland, pituitary gland, lacrimal gland, aorta tree, cerebellum, deep gray matter, gray matter, white matter, external cerebrospinal fluid
unseen lesion	hepatic tumor, breast tumor

B Additional Model Details

We illustrate a detailed version of model architecture of SAM-Med3D in Figure 1. The SAM-Med3D utilizes a holistic 3D structure to capture spatial information directly. In the 3D Image Encoder, patches are firstly embedded using a 3D convolution with a kernel size of (16, 16, 16) and paired with a learnable

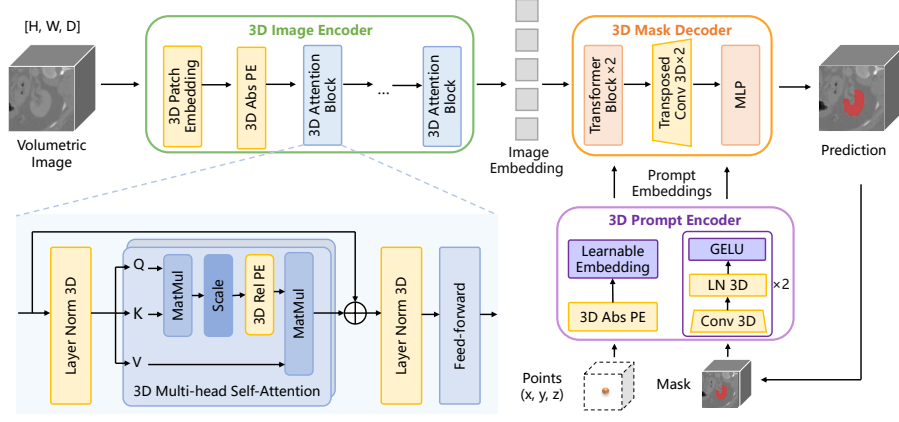


Fig. 1: The holistic 3D structure of our SAM-Med3D. The original 2D components are transformed into their 3D counterparts, encompassing a 3D image encoder, 3D prompt encoder, and 3D mask decoder. 3D convolution, 3D positional encoding (PE) and 3D layer norm are employed to construct the 3D model.

3D absolute Positional Encoding (PE). This encoding is obtained by naturally extending an additional dimension to SAM’s 2D PE. The embeddings of patches are then input to 3D attention blocks. For the 3D attention block, we incorporate a 3D relative PE into the Multi-Head Self-Attention (MHSA) module of SAM, enabling it to directly capture spatial details. Within the prompt encoder, sparse prompts leverage the 3D position encodings to represent 3D spatial nuances, while dense prompts are handled with 3D convolutions. Likewise, the 3D Mask Decoder is integrated with 3D upscaling procedures, employing 3D transposed convolution. Besides, details for preliminary experiment of the encoder for SAM-Med3D can be found in § C.3.

C Additional Experiment Details

C.1 Two-stage Training Procedure

Most of the training settings (except the epoch/iteration) are kept the same for the two stages, with the major difference being the training data (see § A.1). For more details, please see § C.2.

C.2 Implementation Details

Our method is implemented in PyTorch and trained on 2 NVIDIA Tesla A100 GPUs, each with 80GB memory. We use the Adam optimizer with an initial learning rate of $8e-4$ and train for a total of 1000 epochs (i.e. 800 epochs for the 1st-stage pre-training and 200 epochs for the 2nd-stage fine-tuning). During each 200-epoch cycle, the learning rate was reduced to a tenth of its original

value at the 120th, 160th, and 190th epochs. Subsequently, the learning rate was restored to its initial value at the completion of each 200-epoch cycle. This procedure was iterated for a total of four loops. During training, our patch-based pipeline will crop or pad all images to a resolution of $128 \times 128 \times 128$. This crop-or-pad strategy involves padding the edges with zeros for images with width, height, and depth smaller than 128 while using trilinear interpolation to resize images in other cases. The loss function supervising the mask predictions is a DiceCELoss (Averaged Dice loss and CE loss). During our parallel training, the in-total batch size was set to 12, and the interval for aggregation and update of gradient accumulation strategy was set to every 20 steps, with a weight decay of 0.1. As for the data augmentation, we only use RandomFlip on each medical image data.

C.3 Preliminary Experiment Details

For the preliminary experiment, we trained all the models on the training set of *AMOS* [5] for 400 epochs, with the implementation details consistent with those described in § C.2. The three methods were tested on *Totalseg-Test* and *AMOS-Val*, with settings of 1 and 10 prompt points. The results for both prompt quantities were averaged and presented in the main text. Specifically, the ‘seen’ categories in the main text include all 13 organ types from *AMOS*, while the ‘unseen’ categories are those in *Totalseg-Test* not covered by *AMOS* (excluding vertebrae and ribs, considering the difficulty in generalizing from organs to skeletal structures).

Regarding the implementation of specific methods, the 3D Adapter was based on the approach of 3DSAM-Adapter [3]. We used the image encoder part from the open-source code of 3DSAM-Adapter to construct a 3D adapter based on a frozen SAM, while the prompt encoder and mask decoder used the same architecture as SAM-Med3D. For fine-tuning with SAM pre-training, the 3D model structure was the same as SAM-Med3D but the weights were provided by the original SAM. The core challenge of implementation was how to apply 2D SAM weights to the 3D structure. We reused only the weights of the image encoder of SAM, employing a strategy to replicate 2D weights for estimating 3D weights. Specifically, for the 2D convolution in patch embedding (16×16), we repeated it 16 times to obtain $16 \times 16 \times 16$ 3D weights for the 3D convolution. For all the Positional Encodings, we directly used linear/trilinear interpolation to handle different resolutions and duplicated the weights of the x-axis to initialize the newly added z-axis when needed. Finally, the implementation of training from scratch was a straightforward training of SAM-Med3D.

References

1. Baid, U., Ghodasara, S., Bilello, M., Mohan, S., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., Prevedello, L.M., Rudie, J.D., Sako, C., Shinohara, R.T., Bergquist, T., Chai, R., Eddy, J.A., Elliott, J., Reade,

- W., Schaffter, T., Yu, T., Zheng, J., Annotators, B., Davatzikos, C., Mongan, J., Hess, C., Cha, S., Villanueva-Meyer, J.E., Freymann, J.B., Kirby, J.S., Wiestler, B., Crivellaro, P., Colen, R.R., Kotrotsou, A., Marcus, D.S., Milchenko, M., Nazeri, A., Fathallah-Shaykh, H.M., Wiest, R., Jakab, A., Weber, M., Mahajan, A., Menze, B.H., Flanders, A.E., Bakas, S.: The RSNA-ASNR-MICCAI brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *CoRR abs/2107.02314* (2021), <https://arxiv.org/abs/2107.02314>
2. Crimi, A., Bakas, S., Kuijf, H., Menze, B., Reyes, M.: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers, vol. 10670. Springer (2018)
 3. Gong, S., Zhong, Y., Ma, W., Li, J., Wang, Z., Zhang, J., Heng, P.A., Dou, Q.: 3dsam-adaptor: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation. *arXiv preprint arXiv:2306.13465* (2023)
 4. Heller, N., Isensee, F., Trofimova, D., Tejpaul, R., Zhao, Z., Chen, H., Wang, L., Golts, A., Khapun, D., Shats, D., Shoshan, Y., Gilboa-Solomon, F., George, Y., Yang, X., Zhang, J., Zhang, J., Xia, Y., Wu, M., Liu, Z., Walczak, E., McSweeney, S., Vasdev, R., Hornung, C., Solaiman, R., Schoepfoerster, J., Abernathy, B., Wu, D., Abdulkadir, S., Byun, B., Spriggs, J., Struyk, G., Austin, A., Simpson, B., Hagstrom, M., Virnig, S., French, J., Venkatesh, N., Chan, S., Moore, K., Jacobsen, A., Austin, S., Austin, M., Regmi, S., Papanikolopoulos, N., Weight, C.: The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct (2023)
 5. Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *arXiv preprint arXiv:2206.08023* (2022)
 6. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault—workshop and challenge. In: *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*. vol. 5, p. 12 (2015)
 7. Mendrik, A.M., Vincken, K.L., Kuijf, H.J., Breeuwer, M., Bouvy, W.H., De Bresser, J., Alansary, A., De Bruijne, M., Carass, A., El-Baz, A., et al.: Mrbrains challenge: online evaluation framework for brain image segmentation in 3t mri scans. *Computational intelligence and neuroscience* **2015**, 1–1 (2015)
 8. Payette, K., de Dumast, P., Kebiri, H., Ezhov, I., Paetzold, J.C., Shit, S., Iqbal, A., Khan, R., Kottke, R., Grethen, P., et al.: An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. *Scientific data* **8**(1), 167 (2021)
 9. Podobnik, G., Strojjan, P., Peterlin, P., Ibragimov, B., Vrtovec, T.: Han-seg: The head and neck organ-at-risk ct and mr segmentation dataset. *Medical physics* **50**(3), 1917–1927 (2023)
 10. Quinton, F., Popoff, R., Presles, B., Leclerc, S., Meriaudeau, F., Nodari, G., Lopez, O., Pellegrinelli, J., Chevallier, O., Ginhac, D., et al.: A tumour and liver automatic segmentation (atlas) dataset on contrast-enhanced magnetic resonance imaging for hepatocellular carcinoma. *Data* **8**(5), 79 (2023)
 11. Radl, L., Jin, Y., Pepe, A., Li, J., Gsaxner, C., Zhao, F.h., Egger, J.: Avt: Multicenter aortic vessel tree cta dataset collection with ground truth segmentation masks. *Data in brief* **40**, 107801 (2022)
 12. Sun, Y., Gao, K., Wu, Z., Li, G., Zong, X., Lei, Z., Wei, Y., Ma, J., Yang, X., Feng, X., et al.: Multi-site infant brain segmentation algorithms: the iseg-2019 challenge. *IEEE Transactions on Medical Imaging* **40**(5), 1363–1376 (2021)

13. Sun, Y., Wang, L., Jewells, V., Humphreys, K.L., Lin, W.: MICCAI Grand Challenge on Multi-domain Cross-time- point Infant Cerebellum MRI Segmentation 2022 (Mar 2022). <https://doi.org/10.5281/zenodo.6362381>, <https://doi.org/10.5281/zenodo.6362381>
14. Wang, K.: Tumor detection, segmentation and classification challenge on automated 3d breast ultrasound (abus) 2023 (2023), <https://tdsc-abus2023.grand-challenge.org/TDSC-ABUS2023/>
15. Wang, L., Nie, D., Li, G., Puybureau, É., Dolz, J., Zhang, Q., Wang, F., Xia, J., Wu, Z., Chen, J.W., et al.: Benchmark on automatic six-month-old infant brain segmentation algorithms: the iseg-2017 challenge. *IEEE transactions on medical imaging* **38**(9), 2219–2230 (2019)
16. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence* **5**(5), e230024 (2023)