A Light-weight Universal Medical Segmentation Network for Laptops Based on Knowledge Distillation

Songxiao Yang[1 [0000−0001−9036−4817]], Yizhou Li[1 [0000−0002−7122−2087]], Ye Chen[2 [0009−0009−5564−1976]], Zhuofeng Wu[1 [0009−0005−4690−7403]], and Masatoshi Okutomi[1 [0000−0001−5787−0742]]

[1] Tokyo Institute of Technology, Ookayama 2-12-1, Meguro, Tokyo, Japan
[2] The University of Tokyo, Kashiwanoha 5-1-5, Chiba, Japan
syang@ok.sc.e.titech.ac.jp, yli@ok.sc.e.titech.ac.jp, chenye@g.ecc.u-tokyo.ac.jp, zwu@ok.sc.e.titech.ac.jp, mxo@sc.e.titech.ac.jp

Abstract. In medical imaging, accurate and efficient segmentation is crucial for diagnostics, treatment planning, and monitoring disease progression. Traditional methods, while capable of providing reliable results, often require substantial computational resources, which may not be feasible on devices with limited capabilities such as standard CPUs and limited RAM. To address this challenge, we present an optimized universal segmentation framework that leverages a lightweight image encoder RepViT-M0.6, distilled from Swin-T. Our extensive evaluations of the online validation set demonstrate that our approach outperforms the baseline LiteMedSAM model, showing improvements in the Dice Similarity Coefficient (DSC) by approximately 2% and in the Normalized Surface Dice (NSD) by around 1%. Furthermore, the method achieves a more than threefold increase in inference speed, making it viable for real-time applications on devices with limited computational power. This demonstrates that our adaptation significantly enhances processing speed and resource efficiency without sacrificing accuracy.

Keywords: Segment Anything · Lightweight Model · Medical Imaging Segmentation · Computational Efficiency.

1 Introduction

Segmentation plays a crucial role in medical imaging analysis, involving the identification and delineation of regions of interest (ROI) within medical images. The precision of segmentation is crucial for numerous clinical tasks, including disease diagnosis, treatment planning, and monitoring disease progression [13, 50]. Traditionally, manual segmentation has been regarded as the standard for precisely defining anatomical and pathological regions. However, this method is highly time and labor-consuming and demands significant expertise. To overcome these

limitations, automatic segmentation techniques have been introduced. These advanced methods greatly reduce the required time and effort, improve consistency, and enable the efficient analysis of large-scale medical datasets [63].

Recently, deep-learning techniques for image segmentation have shown promising results by training networks to understand intricate image features and produce accurate segmentations [7]. However, many existing models designed for medical image segmentation face a significant limitation that they are tailored for specific tasks and may not perform well when applied to new tasks or different datasets [47]. This task-specific nature poses a challenge to the widespread use of these models in clinical settings. Conversely, recent advancements in natural image segmentation have introduced foundation models, like the segment anything (SAM) [34] and segment everything everywhere all at once [71], showing exceptional adaptability and performance across a range of segmentation tasks. Moreover, the development of MedSAM [42] aims to address the challenge of limited generalizability in medical image segmentation by facilitating universal segmentation across diverse medical imaging tasks.

Despite their strong performance, these methods often utilize large-scale image encoders, leading to high computational demands that limit their practicality. To address this issue and speed up inference while conserving resources, various approaches have been explored to replace the image encoder of SAM with lightweight models. For instance, MobileSAM [68] distills the knowledge of SAM's ViT-H model into a compact vision transformer, while EdgeSAM [70] employs a CNN-based model trained to mimic ViT-H, incorporating a meticulous distillation strategy with the prompt encoder and mask decoder. Additionally, EfficientSAM [65] leverages the MAE pretraining method to enhance performance. EfficientViT-SAM [69] overcomes this limitation by utilizing EfficientViT to substitute SAM's image encoder. However, these methods typically suffer from significant performance drops.

In our work, we propose a solution to further accelerate inference and reduce resource usage while maintaining high performance. Firstly, we enhance the performance of the original LiteMedSAM by replacing its image encoder with Swin-T. Subsequently, to make the encoder lightweight, we distill a RepViT-M0.6 from Swin-T and substitute the encoder of Swin-T with the distilled RepViT-M0.6 image encoder, achieving higher speed and reduced resource consumption while preserving performance.

We extensively evaluate our proposal on the online validation set and compare it with the baseline model (LiteMedSAM). Our results demonstrate improved performance, with the evaluation metric DSC increasing by approximately 2% and NSD by around 1%. Furthermore, we achieve over three times faster inference speed on devices equipped with a CPU and 8GB of RAM.

2 Method

2.1 Pre-processing

We first conducted a statistical analysis on our dataset. As shown in Table 1, Computed Tomography (CT) is the predominant modality, comprising 76.70% of the dataset with 1,219,765 slices. Magnetic Resonance (MR) images are also well-represented, making up 13.55% with 214,454 slices. Positron Emission Tomography (PET) accounts for 4.03%, contributing 64,163 slices. Endoscopy and X-Ray images constitute smaller portions at 2.82% and 1.91% respectively, providing 44,804 and 30,360 slices. Other modalities such as Ultrasound, Dermoscopy, Optical Coherence Tomography (OCT), Mammography, and Fundus Photography each contribute less than 1%.

During the pre-processing of the external public datasets, we initially excluded all slices or images lacking targets or containing extremely small targets (smaller than 20 pixels) to ensure each slice or image had at least one target for segmentation. Subsequently, we normalized all slices/images to a range of [0, 1] and stored each slice/image along with its corresponding ground truth in a single npy file to facilitate faster I/O operations.

In the training phase, all grayscale images were converted to 3-channel images by replicating the image three times along the channel dimension. We resized the longer side of all images to 256 pixels while maintaining the original aspect ratio and then padded them to 256×256 pixels to meet the input requirements of the encoder. If an image had multiple labels, one label was randomly selected. Random data augmentation was applied to both images and their corresponding ground truths. Additionally, we utilized multiple worker processes to accelerate data loading.

Table 1. Statistical analysis of the dataset.

Modality	Proportion	Num. Slices
CT	76.70%	1219765
MR	13.55%	215454
PET	4.03%	64163
Endoscopy	2.82%	44804
X-Ray	1.91%	30360
US	0.40%	6318
Dermoscopy	0.24%	3874
OCT	0.09%	1436
Mammography	0.08%	1233
Fundus	0.07%	1100
Total		1590134

2.2 Proposed Method

The proposed method employs a 2-stage training protocol for a teacher-student model. In the first stage, we train a strong teacher model by replacing MedSAM's image encoder with a Swin-T-based encoder. In the second stage, we distill the features to a RepViT-M0.6-based MedSAM. The following sections will provide details of our model structures and the training & inference strategies.

Teacher Model As previous studies have shown [65,68,69], the image encoder is the heaviest and most parameter-intensive part of the SAM [34], significantly affecting segmentation performance. Thus, selecting a strong yet efficient image encoder is crucial. The default encoder for SAM is ViT-H [17], known for its strong capabilities. However, training SAM with ViT-H requires 68 hours on 256 A100 GPUs as mentioned in [34], posing a significant challenge for reproduction or improvement.

To address this, we opt for the Swin-T image encoder [39], a small but effective hierarchical Transformer that uses shifted windows to limit self-attention computation to non-overlapping local windows while allowing cross-window connections. This architecture is efficient, modeling at various scales with linear computational complexity relative to image size, thus reducing the computational burden compared to ViT.

We replace MedSAM's original image encoder with Swin-T and train the entire pipeline from scratch. The Swin-T-based MedSAM shows significant improvement over the TinyViT-based lightweight MedSAM provided by the competition. Although Swin-T is the smallest Swin Transformer, it is still not efficient enough for fast inference on a laptop CPU. Therefore, we will use this model as a strong teacher model in the next section and distill its features into a smaller student model for much faster inference on a laptop CPU.

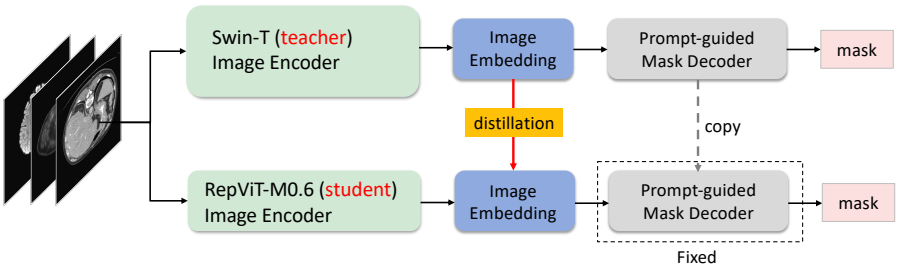

Fig. 1. Proposed teacher-student model architecture. For the teacher model (top), we use a Swin-T image encoder to replace the image encoder in the MedSAM and train the entire pipeline from scratch. For the student model (bottom), which is based on the RepViT-M0.6 image encoder, we distill the features from the teacher image encoder to the student image encoder. The prompt-guided mask decoder is directly copied from the teacher model and not finetuned.

Student Model With a well-trained teacher model, our next step is to select an efficient student model and effectively distill the teacher model's features into it. For the student model, we choose RepViT-M0.6 [62], the smallest version of RepViT, as the image encoder for the student MedSAM model. RepViT is a series of lightweight CNNs redesigned from a ViT perspective, emphasizing their suitability for mobile devices. It builds on the mobile-friendly design of MobileNetV3 [28] and incorporates efficient architectural features of lightweight ViTs.

RepViT, being purely CNN-based, achieves very low latency and memory usage without the computational burden of attentions. After testing various RepViT variants, we found that RepViT-M0.6 offers sufficient performance for feature distillation with the highest inference speed on a CPU.

Feature Distillation to Student Model Next, we discuss the feature distillation from the teacher model to the student model. Following the practice in [68], SAM model distillation methods are classified into fully-coupled, semi-coupled, and decoupled distillation. The first two methods add supervision to the model's final output, i.e., the mask output, while decoupled distillation only distills the image encoder part.

Since the performance bottleneck mainly depends on the image encoder, it is reasonable to fix the prompt-guided mask decoder, which has a small number of parameters, and only distill the image encoder from the feature level. Therefore, we follow this practice and distill the image encoder part, as shown in Fig. 1, using a simple MSE loss between the outputs of the Swin-T encoder and the RepViT encoder. This simple distillation method works surprisingly well, with the student model's performance being comparable to the teacher model.

As mentioned in [68], finetuning the prompt-guided mask decoder after distilling the image encoder might potentially improve overall performance. However, in our case, the small image encoder with RepViT-M0.6 sufficiently matches the Swin-T in the feature level. Thus, finetuning the prompt-guided mask decoder with mask loss did not provide a performance boost.

Loss functions. For the teacher MedSAM model with Swin-T, we train the entire pipeline from scratch using a combination of Dice loss and focal loss. This compound loss function is robust for various medical image segmentation tasks [41]. For the student MedSAM model with RepViT-M0.6, we distill only the image encoder part. We compute the MSE loss between the feature outputs of the teacher and student models' image encoders.

Strategies to Accelerate CPU Inference Our student model with RepViT-M0.6 is already fast on CPU inference. However, we explored quantization for potential benefits. We tried Pytorch FX Graph Quantization [6] and ONNX Runtime [15] for Int8 quantization, but observed significant performance loss even after calibration. Therefore, we abandoned quantization. Instead, we used

"torch.jit" to increase the model loading speed, contributing to the speed boost in the Docker test.

2.3 Post-processing

First, as the model outputs are logits, we first convert the logits to probabilities using a sigmoid function. The masks are then cropped to match the new size, which corresponds to the image shape after resizing to the longest side of 256 pixels. Following this, the masks are resized back to the original image dimensions using bilinear interpolation, ensuring proper alignment and smooth transitions. The resulting tensor is then converted to a NumPy array on the CPU. Finally, a threshold is applied to generate a binary segmentation mask, where values greater than 0.5 indicate the presence of the target object. This comprehensive process ensures that the model's raw outputs are accurately transformed into a practical segmentation format.

3 Experiments

3.1 Dataset and evaluation measures

We use the challenge dataset and external public datasets for network training, as shown in the supplementary Table 9.

The evaluation metrics include two accuracy measures—Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD)—and the running time for efficiency measurement. These metrics collectively contribute to the ranking computation.

3.2 Environment settings

The details of our environment are presented in Table 2. We use Ubuntu 22.04.4 LTS as our operating system. Our system is equipped with an Intel(R) Core(TM) i9-13900KF CPU and 64GB of RAM. Additionally, we utilize an NVIDIA RTX 4090 GPU with 24GB of memory.

3.3 Training protocols of LiteMedSAM with Swin-T image encoder

In training LiteMedSAM with the Swin-T image encoder, we initially apply data augmentation techniques to enhance model robustness. These techniques include random horizontal and vertical flips. To avoid overfitting to the data sequence, we randomly select images from the dataset. For images with multiple labels, we randomly choose one label per image. A validation set is constructed by randomly selecting approximately 5% of the entire training dataset.

As shown in Table 3, during training, images are pre-processed to $3 \times 256 \times 256$. The network is trained from scratch over 100 epochs. We employ a combination of Dice Loss, Cross Entropy Loss, and Mean Squared Error Loss (MSELoss)

Table 2. Development environments and requirements.

System	Ubuntu 22.04.4 LTS
CPU	Intel(R) Core(TM) i9-13900KF CPU@3.00GHz
RAM	16×4GB; 2.67MT/s
GPU (number and type)	One NVIDIA RTX 4090 24G
CUDA version	12.1
Programming language	Python 3.10
Deep learning framework	torch 2.1.2, torchvision 0.16.2
Specific dependencies	N/A
Code	GitHub

as the loss function. The initial learning rate is set to 0.005. We use AdamW as the optimizer and ReduceLROnPlateau as the learning rate scheduler, which reduces the learning rate by a factor of 0.9 whenever the validation loss does not decrease for five consecutive epochs. We assess the model's performance on the validation set at the end of each epoch and save the model that records the best performance on the validation set.

Table 3. Training protocols of LiteMedSAM with Swin-T image encoder.

Pre-trained Model	N/A
Batch size	8
Patch size	3×256×256
Total epochs	100
Optimizer	AdamW
Initial learning rate (lr)	0.005
Lr decay schedule	ReduceLROnPlateau(reduction ratio 0.9)
Training time	1200 hours
Loss function	Dice Loss, Cross Entropy Loss, MSE Loss
Number of model parameters	14.55M
Number of flops	42.85G
CO_2eq	848 Kg

3.4 Training protocols for the knowledge distillation of RepViT-M0.6 image encoder from Swin-T image encoder

For the knowledge distillation of the RepViT-M0.6 image encoder from the Swin-T image encoder, the training process is similar to the training of LiteMedSAM with Swin-T. The data is processed by data augmentation and shuffled before input into the network. We maintain the same dataset split as used in the training of LiteMedSAM with Swin-T. During training, images are pre-processed to $3 \times$

256 × 256 and we conduct the distillation of the RepViT-M0.6 from the Swin-T image encoder over 50 epochs, as illustrated in Table 4. To minimize the difference in the image embedding outputs between RepViT-M0.6 and Swin-T, we calculate MSELoss. We utilize AdamW as the optimizer with an initial learning rate of 0.005 and ReduceLROnPlateau as the learning rate scheduler, which reduces the learning rate by a factor of 0.9 whenever the validation loss does not decrease for five epochs. We evaluate the model on the validation set after each epoch and save the model version that achieved the lowest validation loss.

Table 4. Training protocols for the knowledge distillation of RepViT-M0.6 image encoder from Swin-T image encoder.

Pre-trained Teacher Model	Swin-T Image Encoder
Pre-trained Student Model	N/A
Batch size	8
Patch size	3×256×256
Total epochs	50
Optimizer	AdamW
Initial learning rate (lr)	0.005
Lr decay schedule	ReduceLROnPlateau(reduction ratio 0.9)
Training time	400 hours
Loss function	MSE Loss
Number of model parameters	2.32M
Number of flops	9.00G
CO_2eq	99 Kg

4 Results and discussion

4.1 Quantitative results on online validation set

In Table 5, we compare three methods: the baseline, LiteMedSAM with Swin-T, and our proposed LiteMedSAM with RepViT-M0.6 image encoder, which is distilled from the Swin-T model. We evaluate their performance on the online validation set using the DSC and NSD evaluation metrics.

The LiteMedSAM (without knowledge distillation) demonstrates an average improvement of approximately 2% in both DSC and NSD compared to the baseline. This improvement is observed across most modalities, with the exception of a slight decrease in Endoscopy and Fundus. Furthermore, when employing knowledge distillation, there is only a minor decline in the average DSC and NSD, yet still shows a clear improvement compared to the baseline.

Table 5. Quantitative evaluation results on online validation set.

Target	Baseline		w/o Knowledge Distillation		Proposed	
	DSC(%)	NSD(%)	DSC(%)	NSD(%)	DSC(%)	NSD (%)
CT	89.53	91.82	89.94	91.85	**92.65**	**95.06**
MR	78.75	81.87	81.35	84.36	**86.09**	**89.34**
PET	68.91	55.43	**71.00**	**55.95**	62.38	38.58
US	81.34	87.12	81.60	86.74	**82.45**	**87.54**
X-Ray	70.23	76.58	78.39	**84.49**	**79.13**	85.11
Dermoscopy	92.65	94.14	**93.58**	**95.08**	93.45	94.96
Endoscopy	**94.87**	**97.38**	93.87	96.43	93.48	96.23
Fundus	**95.85**	**97.48**	95.47	97.11	94.68	96.36
Microscopy	71.79	76.95	77.27	83.88	**77.80**	**84.38**
Average	82.66	84.31	**84.72**	**86.21**	84.68	85.28

Table 6. Quantitative evaluation of segmentation efficiency in terms of running time (s).

Case ID	Size	Num. Objects	Baseline	w/o Knowledge Distillation	Proposed
3DBox_CT_0566	(287, 512, 512)	6	206.4344	212.4705	**48.5365**
3DBox_CT_0888	(237, 512, 512)	6	55.7089	55.686	**13.6685**
3DBox_CT_0860	(246, 512, 512)	1	7.7789	7.6037	**2.502**
3DBox_MR_0621	(115, 400, 400)	6	101.7202	89.6444	**20.3509**
3DBox_MR_0121	(64, 290, 320)	6	58.291	50.8915	**13.5169**
3DBox_MR_0179	(84, 512, 512)	1	8.1488	7.0312	**1.9713**
3DBox_PET_0001	(264, 200, 200)	1	5.511	3.8106	**1.3485**
2DBox_US_0525	(256, 256, 3)	1	0.4136	0.4332	**0.1382**
2DBox_X-Ray_0053	(320, 640, 3)	34	1.3110	1.3009	**1.271**
2DBox_Dermoscopy_0003	(3024, 4032, 3)	1	0.7331	0.6091	**0.4595**
2DBox_Endoscopy_0086	(480, 560, 3)	1	0.4311	0.4164	**0.1533**
2DBox_Fundus_0003	(2048, 2048, 3)	1	0.4785	0.3656	**0.1998**
2DBox_Microscope_0008	(1536, 2040, 3)	19	0.9869	0.9626	**0.6367**
2DBox_Microscope_0016	(1920, 2560, 3)	241	8.7242	8.6334	**8.1791**

4.2 Qualitative results on online validation set

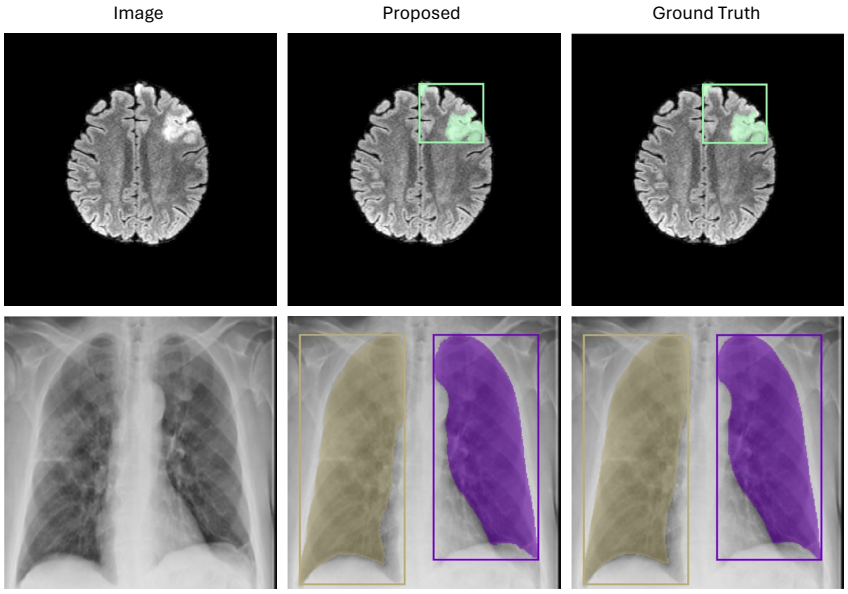

Fig. 2. The examples of good segmentation results.

In this section, we present examples of both good and bad segmentation results to further analyze the performance of our proposed method. Generally, our method performs well with images that exhibit high contrast, simple backgrounds, and regular shapes. However, it encounters challenges with images that contain overlapping textures and irregular shapes and borders.

As illustrated in Fig 2, in the first row, the brain MR image showcases high contrast against a simple background, while in the second row, the chest X-ray displays the lung in regular shapes, which our model can accurately segment. Conversely, as depicted in Fig 3, the X-ray image exhibits overlapping textures, and the targets within the box are irregularly shaped in both the first and second rows, posing difficulties for accurate segmentation with our method.

4.3 Segmentation efficiency results on online validation set

An important challenge for this task is the constraint of the target device, which is equipped with only a CPU and limited memory (8GB RAM), making segmentation efficiency crucial. We present some challenging cases that require longer processing times in Table 6. Our proposed method consistently demonstrates a significant reduction in running time across almost all cases compared to both the baseline and the LiteMedSAM with Swin-T. This improvement is particularly

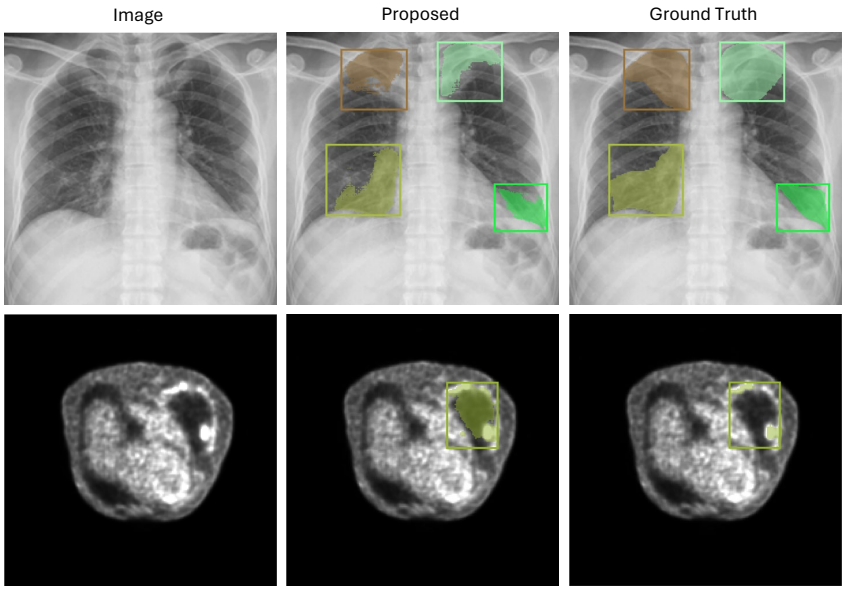

Fig. 3. The examples of bad segmentation results.

notable in complex 3D imaging cases. For instance, in the 3DBox_CT_0566 case, the proposed method reduces the running time from over 200 seconds in the baseline to just 48.54 seconds. Similarly, in the 3DBox_CT_0888 and 3DBox_CT_0860 cases, the running times decrease from 55.71 and 7.78 seconds in the baseline to 13.67 and 2.50 seconds, respectively.

In 2D imaging scenarios, such as 2DBox_US_0525 and 2DBox_X-Ray_0053, the proposed method drastically reduces running times to 0.14 and 1.27 seconds from 0.41 and 1.31 seconds in the baseline, respectively. The reductions are even more striking in cases like 2DBox_Dermoscopy_0003 and 2DBox_Endoscopy_0086, where the proposed method achieves running times of 0.46 and 0.15 seconds, down from 0.73 and 0.43 seconds in the baseline. These results underscore the effectiveness of our proposed method in enhancing the efficiency of segmentation on devices with limited computational resources.

4.4 Results on final testing set

4.5 Limitation and future work

Although our proposed method has exhibited promising performance and excellent efficiency, it still encounters difficulties with cases characterized by low contrast, irregular shapes, and overlapping textures, as mentioned in Section 4.2. These challenges highlight areas for improvement in future work.

One potential avenue for addressing these issues could involve training a more robust teacher model to provide better guidance during knowledge distillation.

Additionally, incorporating more diverse and challenging data into the training process could help the model learn to handle such cases more effectively.

5 Conclusion

In this study, we introduced an optimized segmentation framework leveraging distillation techniques to enhance the efficiency of medical segmentation networks. By integrating distillation knowledge, notable reductions in inference time were achieved while preserving segmentation accuracy, thus demonstrating promising prospects for real-time application in resource-constrained environments. Specifically, our results demonstrate a significant threefold decrease in processing time compared to the baseline model, along with improvements in quantitative evaluation.

Despite these advancements, challenges remain with low-contrast images, irregular shapes, and overlapping textures. Future work will aim to address these issues, enhancing the robustness and applicability of our proposed framework.

Acknowledgements We thank all the data owners for making the medical images publicly available and CodaLab [66] for hosting the challenge platform.

References

1. Aerts, H.J., Velazquez, E.R., Leijenaar, R.T., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., et al.: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature communications **5**(1), 4006 (2014)
2. Ahmed, A., Elmohr, M., Fuentes, D., Habra, M., Fisher, S., Perrier, N., Zhang, M., Elsayes, K.: Radiomic mapping model for prediction of ki-67 expression in adrenocortical carcinoma. Clinical Radiology **75**(6), 479–e17 (2020)
3. Ahmed, Z., Panhwar, S.Q., Baqai, A., Umrani, F.A., Ahmed, M., Khan, A.: Deep learning based automated detection of intraretinal cystoid fluid. International Journal of Imaging Systems and Technology **32**(3), 902–917 (2022)
4. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
5. An, N.S., Lan, P.N., Hang, D.V., Long, D.V., Trung, T.Q., Thuy, N.T., Sang, D.V.: Blazeneo: Blazing fast polyp segmentation and neoplasm detection. IEEE Access **10**, 43669–43684 (2022). https://doi.org/10.1109/ACCESS.2022.3168693
6. Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., et al.: Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In: Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2. pp. 929–947 (2024)
7. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)

8. Bakr, S., Gevaert, O., Echegaray, S., Ayers, K., Zhou, M., Shafiq, M., Zheng, H., Benson, J.A., Zhang, W., Leung, A.N., et al.: A radiogenomic dataset of non-small cell lung cancer. Scientific data **5**(1), 1–9 (2018)
9. Bowen, S.R., Yuh, W.T., Hippe, D.S., Wu, W., Partridge, S.C., Elias, S., Jia, G., Huang, Z., Sandison, G.A., Nelson, D., et al.: Tumor radiomic heterogeneity: Multiparametric functional imaging to characterize variability and predict response following cervical cancer radiation therapy. Journal of Magnetic Resonance Imaging **47**(5), 1388–1396 (2018)
10. Campello, V.M., Gkontra, P., Izquierdo, C., Martín-Isla, C., Sojoudi, A., Full, P.M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., Parreño, M., Albiol, A., Kong, F., Shadden, S.C., Acero, J.C., Sundaresan, V., Saber, M., Elattar, M., Li, H., Menze, B., Khader, F., Haarburger, C., Scannell, C.M., Veta, M., Carscadden, A., Punithakumar, K., Liu, X., Tsaftaris, S.A., Huang, X., Yang, X., Li, L., Zhuang, X., Viladés, D., Descalzo, M.L., Guala, A., Mura, L.L., Friedrich, M.G., Garg, R., Lebel, J., Henriques, F., Karakas, M., Çavuş, E., Petersen, S.E., Escalera, S., Seguí, S., Rodríguez-Palomares, J.F., Lekadir, K.: Multi-centre, multi-vendor and multi-disease cardiac segmentation: The mms challenge. IEEE Transactions on Medical Imaging **40**(12), 3543–3554 (2021). https://doi.org/10.1109/TMI.2021.3090082
11. Cheng, J., Huang, W., Cao, S., Yang, R., Yang, W., Yun, Z., Wang, Z., Feng, Q.: Enhanced performance of brain tumor classification via tumor region augmentation and partition. PloS one **10**(10), e0140381 (2015)
12. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al.: The cancer imaging archive (tcia): maintaining and operating a public information repository. Journal of digital imaging **26**, 1045–1057 (2013)
13. De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., et al.: Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature medicine **24**(9), 1342–1350 (2018)
14. Degerli, A., Kiranyaz, S., Hamid, T., Mazhar, R., Gabbouj, M.: Early myocardial infarction detection over multi-view echocardiography. Biomedical Signal Processing and Control **87**, 105448 (2024)
15. developers, O.R.: Onnx runtime. https://onnxruntime.ai/ (2021), version: x.y.z
16. Dorent, R., Kujawa, A., Ivory, M., Bakas, S., Rieke, N., Joutard, S., Glocker, B., Cardoso, J., Modat, M., Batmanghelich, K., et al.: Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation. Medical Image Analysis **83**, 102628 (2023)
17. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
18. Fedorov, A., Schwier, M., Clunie, D., Herz, C., Pieper, S., Kikinis, R., Tempany, C., Fennessy, F.: An annotated test-retest collection of prostate multiparametric mri. Scientific data **5**(1), 1–13 (2018)
19. Gao, S., Zhou, H., Gao, Y., Zhuang, X.: Bayeseg: Bayesian modeling for medical image segmentation with interpretable generalizability. Medical Image Analysis **89**, 102889 (2023)
20. Gatidis, S., Hepp, T., Früh, M., La Fougère, C., Nikolaou, K., Pfannenberg, C., Schölkopf, B., Küstner, T., Cyran, C., Rubin, D.: A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. Scientific Data **9**(1), 601 (2022)

21. Glaister, J., Wong, A., Clausi, D.A.: Automatic segmentation of skin lesions from dermatological photographs using a joint probabilistic texture distinctiveness approach. IEEE Transactions on Biomedical Engineering (2014)
22. Hatamizadeh, A.: An Artificial Intelligence Framework for the Automated Segmentation and Quantitative Analysis of Retinal Vasculature. University of California, Los Angeles (2020)
23. Hatamizadeh, A., Hosseini, H., Patel, N., Choi, J., Pole, C.C., Hoeferlin, C.M., Schwartz, S.D., Terzopoulos, D.: Ravir: A dataset and methodology for the semantic segmentation and quantitative analysis of retinal arteries and veins in infrared reflectance imaging. IEEE Journal of Biomedical and Health Informatics **26**(7), 3272–3283 (2022)
24. Heller, N., Isensee, F., Maier-Hein, K.H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., 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 **67**, 101821 (2021)
25. Hernandez Petzsche, M.R., de la Rosa, E., Hanning, U., Wiest, R., Valenzuela, W., Reyes, M., Meyer, M., Liew, S.L., Kofler, F., Ezhov, I., et al.: Isles 2022: A multicenter magnetic resonance imaging stroke lesion segmentation dataset. Scientific data **9**(1), 762 (2022)
26. van den Heuvel, T.L., de Bruijn, D., de Korte, C.L., Ginneken, B.v.: Automated measurement of fetal head circumference using 2d ultrasound images. PloS one **13**(8), e0200412 (2018)
27. Hong, W.Y., Kao, C.L., Kuo, Y.H., Wang, J.R., Chang, W.L., Shih, C.S.: Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. arXiv preprint arXiv:2012.12453 (2020)
28. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1314–1324 (2019)
29. Jaeger, S., Karargyris, A., Candemir, S., Folio, L., Siegelman, J., Callaghan, F., Xue, Z., Palaniappan, K., Singh, R.K., Antani, S., et al.: Automatic tuberculosis screening using chest radiographs. IEEE transactions on medical imaging **33**(2), 233–245 (2013)
30. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., De Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26. pp. 451–462. Springer (2020)
31. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in Neural Information Processing Systems **35**, 36722–36732 (2022)
32. Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., et al.: Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. Medical Image Analysis **69**, 101950 (2021)
33. Khaled, R., Helal, M., Alfarghaly, O., Mokhtar, O., Elkorany, A., El Kassas, H., Fahmy, A.: Categorized contrast enhanced mammography dataset for diagnostic and artificial intelligence research. Scientific Data **9**(1), 122 (2022)
34. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of the International Conference on Computer Vision. pp. 4015–4026 (2023)

35. Kiser, K.J., Barman, A., Stieb, S., Fuller, C.D., Giancardo, L.: Novel autosegmentation spatial similarity metrics capture the time required to correct segmentations better than traditional metrics in a thoracic cavity segmentation workflow. Journal of Digital Imaging **34**, 541–553 (2021)
36. Kovalyk, O., Morales-Sánchez, J., Verdú-Monedero, R., Sellés-Navarro, I., Palazón-Cabanes, A., Sancho-Gómez, J.L.: Papila: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. Scientific Data **9**(1), 291 (2022)
37. Kuijf, H.J., Biesbroek, J.M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M.J., Casamitjana, A., 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)
38. Litjens, G., Toth, R., Van De Ven, W., Hoeks, C., Kerkstra, S., Van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., et al.: Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. Medical image analysis **18**(2), 359–373 (2014)
39. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
40. Lu, Y., Zhou, M., Zhi, D., Zhou, M., Jiang, X., Qiu, R., Ou, Z., Wang, H., Qiu, D., Zhong, M., et al.: The jnu-ifm dataset for segmenting pubic symphysis-fetal head. Data in Brief **41**, 107904 (2022)
41. Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X., Martel, A.L.: Loss odyssey in medical image segmentation. Medical Image Analysis **71**, 102035 (2021)
42. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
43. Ma, J., Wang, Y., An, X., Ge, C., Yu, Z., Chen, J., Zhu, Q., Dong, G., He, J., He, Z., et al.: Toward data-efficient learning: A benchmark for covid-19 ct lung and infection segmentation. Medical physics **48**(3), 1197–1210 (2021)
44. Ma, J., Xie, R., Ayyadhury, S., Ge, C., Gupta, A., Gupta, R., Gu, S., Zhang, Y., Lee, G., Kim, J., et al.: The multimodality cell segmentation challenge: toward universal solutions. Nature methods pp. 1–11 (2024)
45. Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., et al.: Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6695–6714 (2021)
46. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
47. Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D.: Image segmentation using deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence **44**(7), 3523–3542 (2021)
48. Morshid, A., Elsayes, K.M., Khalaf, A.M., Elmohr, M.M., Yu, J., Kaseb, A.O., Hassan, M., Mahvash, A., Wang, Z., Hazle, J.D., et al.: A machine learning model to predict hepatocellular carcinoma response to transcatheter arterial chemoembolization. Radiology: Artificial Intelligence **1**(5), e180021 (2019)

49. Myronenko, A., Yang, D., He, Y., Xu, D.: Aorta segmentation from 3d ct in miccai seg. a. 2023 challenge. In: MICCAI Challenge on Segmentation of the Aorta, pp. 13–18. Springer (2023)
50. Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C.P., Heidenreich, P.A., Harrington, R.A., Liang, D.H., Ashley, E.A., et al.: Video-based ai for beat-to-beat assessment of cardiac function. Nature **580**(7802), 252–256 (2020)
51. Payette, K., de Dumast, P., Kebiri, H., Ezhov, I., Paetzold, J.C., Shit, S., Iqbal, A., Khan, R., Kottke, R., Grehten, P., et al.: An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. Scientific data **8**(1), 167 (2021)
52. Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Meriaudeau, F.: Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. Data **3**(3), 25 (2018)
53. Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S.B.A., Islam, M.T., Al Maadeed, S., Zughaier, S.M., Khan, M.S., et al.: Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. Computers in biology and medicine **132**, 104319 (2021)
54. Rister, B., Yi, D., Shivakumar, K., Nobashi, T., Rubin, D.L.: Ct-org, a new dataset for multiple organ segmentation in computed tomography. Scientific Data **7**(1), 381 (2020)
55. Roth, H.R., Lu, L., Seff, A., Cherry, K.M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., Summers, R.M.: A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part I 17. pp. 520–527. Springer (2014)
56. Roth, H.R., Xu, Z., Tor-Díez, C., Jacob, R.S., Zember, J., Molto, J., Li, W., Xu, S., Turkbey, B., Turkbey, E., et al.: Rapid artificial intelligence solutions in a pandemic—the covid-19-20 lung ct lesion segmentation challenge. Medical image analysis **82**, 102605 (2022)
57. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019)
58. Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., et al.: Gland segmentation in colon histology images: The glas challenge contest. Medical image analysis **35**, 489–502 (2017)
59. Song, Y., Zheng, J., Lei, L., Ni, Z., Zhao, B., Hu, Y.: Ct2us: Cross-modal transfer learning for kidney segmentation in ultrasound images with synthesized data. Ultrasonics **122**, 106706 (2022). https://doi.org/https://doi.org/10.1016/j.ultras.2022.106706, https://www.sciencedirect.com/science/article/pii/S0041624X22000191
60. Tahir, A.M., Chowdhury, M.E., Khandakar, A., Rahman, T., Qiblawey, Y., Khurshid, U., Kiranyaz, S., Ibtehaz, N., Rahman, M.S., Al-Maadeed, S., et al.: Covid-19 infection localization and severity grading from chest x-ray images. Computers in biology and medicine **139**, 105002 (2021)
61. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018)

62. Wang, A., Chen, H., Lin, Z., Pu, H., Ding, G.: Repvit: Revisiting mobile cnn from vit perspective. arXiv preprint arXiv:2307.09283 (2023)
63. Wang, G., Zuluaga, M.A., Li, W., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., et al.: Deepigeos: a deep interactive geodesic framework for medical image segmentation. IEEE transactions on pattern analysis and machine intelligence **41**(7), 1559–1572 (2018)
64. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5) (2023)
65. Xiong, Y., Varadarajan, B., Wu, L., Xiang, X., Xiao, F., Zhu, C., Dai, X., Wang, D., Sun, F., Iandola, F., et al.: Efficientsam: Leveraged masked image pretraining for efficient segment anything. arXiv preprint arXiv:2312.00863 (2023)
66. Xu, Z., Escalera, S., Pavão, A., Richard, M., Tu, W.W., Yao, Q., Zhao, H., Guyon, I.: Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. Patterns **3**(7), 100543 (2022)
67. Yang, J., Veeraraghavan, H., Armato III, S.G., Farahani, K., Kirby, J.S., Kalpathy-Kramer, J., van Elmpt, W., Dekker, A., Han, X., Feng, X., et al.: Autosegmentation for thoracic radiation treatment planning: a grand challenge at aapm 2017. Medical physics **45**(10), 4568–4581 (2018)
68. Zhang, C., Han, D., Qiao, Y., Kim, J.U., Bae, S.H., Lee, S., Hong, C.S.: Faster segment anything: Towards lightweight sam for mobile applications. arXiv preprint arXiv:2306.14289 (2023)
69. Zhang, Z., Cai, H., Han, S.: Efficientvit-sam: Accelerated segment anything model without performance loss. In: CVPR Workshop: Efficient Large Vision Models (2024)
70. Zhou, C., Li, X., Loy, C.C., Dai, B.: Edgesam: Prompt-in-the-loop distillation for on-device deployment of sam. arXiv preprint arXiv:2312.06660 (2023)
71. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. Advances in Neural Information Processing Systems **36** (2024)

Table 7. Checklist Table. Please fill out this checklist table in the answer column.

Requirements	Answer
A meaningful title	Yes
The number of authors (≤6)	5
Author affiliations and ORCID	Yes
Corresponding author email is presented	Yes
Validation scores are presented in the abstract	Yes
Introduction includes at least three parts: background, related work, and motivation	Yes
A pipeline/network figure is provided	Figure 1
Pre-processing	Page 3
Strategies to data augmentation	Page 7
Strategies to improve model inference	Page 5
Post-processing	Page 6
Environment setting table is provided	Table 6
Training protocol table is provided	Table 7
Ablation study	Page 9
Efficiency evaluation results are provided	Table 6
Visualized segmentation example is provided	Figure 2 3
Limitation and future work are presented	Yes
Reference format is consistent.	Yes
Main text >= 8 pages (not include references and appendix)	Yes

Table 8. The challenge datasets we used for training.

Dataset	Anatomy	Modality
COVID-19-20 [56]	Chest	CT
AbdomenCT-1K [45]	Abdomen	CT
FDG-PET-CT-Lesions [20]	Whole body	CT
NSCLC Radiogenomics [8]	Chest	CT
NSCLC-Radiomics [1]	Lung	CT
CT Lymph Nodes [55]	Abdomen, Mediastinum	CT
NSCLC-PleuralEffusion [35]	Chest	CT
NSCLC-LungMSD-LUNG [7]	Chest	CT
KiTS23 [24]	Kidney	CT
CT-ORG [54]	whole body	CT
COVID-19-20-CTSEG [43]	Chest	CT
TotalSegmentator [64]	whole body	CT
AMOS [31]	Abdomen	CT, MR
LCTSC [67]	Chest	CT
HCC-TACE-Seg [48]	Liver	CT
Adrenal-ACC-Ki67-Seg [2]	Abdomen	CT
MSD [7]	various	CT, MR
ISLES [25]	Brain	MR
WMH [37]	Brain	MR
BraTS [46]	Head	MR
PROMISE12 [38]	Prostate	MR
MSD-Prostate [57]	Prostate	MR
NCI-ISBI [12]	Prostate	MR
Crossmoda [16]	Brain	MR
QIN-PROSTATE-Repeatability [18]	Prostate	MRI
CC-Tumor Heterogeneity [9]	Cervical Cancer	MR
COVID-19 Radiography Database [53]	Lung	CXR
COVID-QU-Ex [60]	Lung	CXR
Chest Xray Masks and Labels [29]	Chest	CXR
Chest X-Ray Images with Pneumothorax Masks	Chest	CXR
CDD-CESM [33]	Breast	Mammography
Intraretinal Cystoid Fluid [3]	Eye	OCT
ps-fh-aop-2023 [40]	Head	US
hc18 [26]	Fetal Head	US
Breast Ultrasound Images Dataset [4]	Breast	US
ISIC2018 [61]	Skin	Dermoscopy
CholecSeg8k [27]	Abdominal	Endoscopy
Kvasir-SEG [30]	Abdominal	Endoscopy
m2caiSeg	Abdominal	Endoscopy
PAPILA [36]	Eye	Fundus
IDRiD [52]	Eye	Fundus
NeurIPS CellSeg [44]	Cells	Microscopy

Table 9. The external public datasets we used for training.

Dataset	Anatomy	Modality
CT Lung & Hearth & Trachea segmentation	Chest	CT
Seg.A. [49]	Aorta	CT
Figshare Brain Tumor Dataset [11]	Brain	MR
Uwaterloo skin cancer [21]	Skin	Dermoscopy
BKAI-IGH NeoPolyp [5]	Abdominal	Endoscopy
CT2USforKidneySeg [59]	Kidney	US
Ultrasound Nerve Segmentation	Neck	US
GlaS@MICCAI'2015: Gland Segmentation [58]	ColonGland	Microscopy
MM-WHS [19]	Heart	CT, MR
MMs-20-21 [10]	Heart	MR
FeTA [51]	Brain	MR
CHAOS [32]	Abdomen	CT, MR
Drive	Eye	Fundus
RAVIR [22, 23]	Eye	Fundus
FetoPlac	Fetus	Microscopy
HMC-QU [14]	Heart	US