Improved Baselines with Synchronized Encoding for Universal Medical Image Segmentation

Sihan Yang[1], Jiadong Feng[1*], Xuande Mi[1*], Haixia Bi[2(✉)], Hai Zhang[3,5], and Jian Sun[4,5]

[1] Xi'an Jiaotong University, Xi'an, China
[2] School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an, China
haixia.bi@xjtu.edu.cn
[3] School of Mathematics, Northwest University, Xi'an, China
[4] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China
[5] Pazhou Laboratory (Huangpu), Guangzhou, China

Abstract. Large foundation models, known for their strong zero-shot generalization capabilities, can be applied to a wide range of downstream tasks. However, developing foundation models for medical image segmentation poses a significant challenge due to the domain gap between natural and medical images. While fine-tuning techniques based on the Segment Anything Model (SAM) have been explored, they primarily focus on scaling up data or refining inference strategies without incorporating domain-specific architectural designs, limiting their zero-shot performance. To optimize segmentation performance under standard inference settings and provide a strong baseline for future research, we introduce SyncSAM, which employs a synchronized dual-branch encoder that integrates convolution and Transformer features in a synchronized manner to enhance medical image encoding, and a multi-scale dual-branch decoder to preserve image details. SyncSAM is trained on two of the largest medical image segmentation datasets, SA-Med2D-20M and IMed-361M, resulting in a series of pre-trained models for universal medical image segmentation. Experimental results demonstrate that SyncSAM not only achieves state-of-the-art performance on test sets but also exhibits strong zero-shot capabilities on unseen datasets. Code and checkpoints are available at https://github.com/Hhankyangg/SyncSAM.

Keywords: Foundation Models · Medical Image Segmentation · Segment Anything Model.

1 Introduction

Large foundation models pre-trained on extensive datasets provide strong zero-shot capabilities, making them effective for diverse downstream tasks [2,17,3,4].

[*] Equal contributions.

A key example is the Segment Anything Model (SAM) [13], an interactive segmentation model trained on large datasets. Leveraging its robust zero-shot performance, SAM has not only established a stable foundation for segmentation tasks [20] but also advanced various visual and multimodal applications [15,1].

Medical image segmentation is a fundamental task in medical imaging analysis, playing a crucial role in disease diagnosis, treatment planning, and disease progression monitoring [8]. With the emergence of SAM, a corpus of researchers have explored its potential in medical image segmentation. These works can be categorized based on the dataset scale: leveraging SAM for specialist models on small datasets and for foundation models on large-scale datasets. For specialist models trained on small datasets, approaches include Parameter-Efficient Fine-Tuning (PEFT) [27] and architectural modifications such as hierarchical decoders [7] and prompt-decoupled mask decoders [10]. While effective on small datasets, their performance on large datasets remains unverified. For foundation models trained on large-scale datasets, two primary fine-tuning strategies exist. The first fine-tunes SAM at different parameter scales using either PEFT techniques [6] or full-parameter fine-tuning [18], yet these models often lack domain-specific architectural enhancements, leading to suboptimal performance. The second modifies the inference process [30,28] to improve segmentation accuracy, incorporating advanced interactive methods or support-set-based strategies. While effective in complex prompt settings, these modifications may limit broader applicability to downstream tasks [14].

To fill the gap of foundation models in integrating expert-designed modules, we introduce SyncSAM in this work. Inspired by the advantages of Convolutional Neural Network (CNN) in capturing fine-grained features and dealing with noisy images, we incorporate a CNN branch into SAM's image encoder, injecting medical-specific domain bias into the Vision Transformer(ViT) [9] backbone. Moreover, we propose a synchronized fusion strategy that performs stage-wise fusion of ViT and CNN features rather than merging them in a single step. This progressive alignment of local and global representations enhances contextual understanding of medical images. To further improve segmentation, we devise a multi-scale dual-branch decoder that leverages early-stage encoder features to refine fine-grained edge details. By combining the synchronized dual-branch encoder and the multi-scale dual-branch decoder, SyncSAM effectively enhances feature representation, enabling more precise segmentation for medical images.

We train four versions of SyncSAM on two of the largest 2D medical image segmentation datasets, SA-Med2D-20M [29] and IMed-361M [5]. Experimental results on test sets confirm state-of-the-art (SOTA) performance, while zero-shot evaluations on six unseen datasets demonstrate that SyncSAM outperforms all existing medical foundation models using simple box prompts.

This work makes three key contributions. (1) We introduce SyncSAM, a novel foundation model tailored for medical image segmentation, filling the gap in large-scale SAM-based foundation models with expert-designed modules. (2) We train and evaluate SyncSAM on two of the largest 2D medical segmentation datasets, achieving SOTA performance on test sets while demonstrating strong

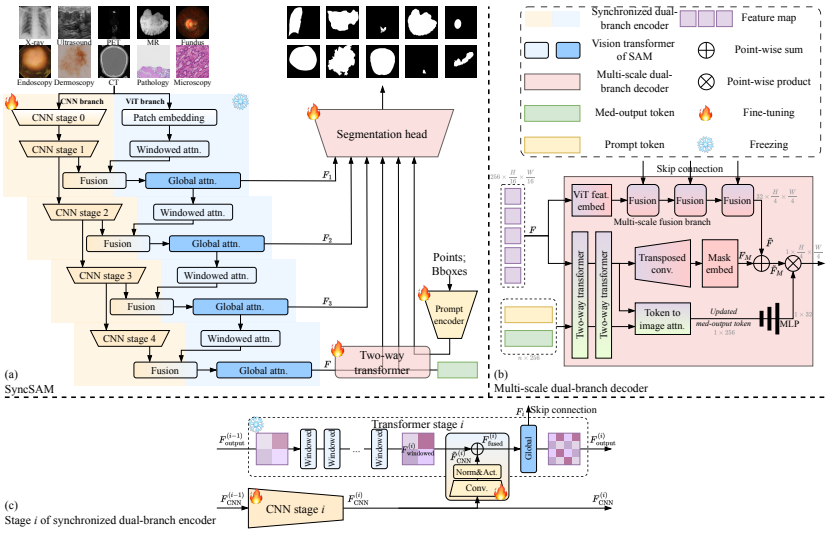

Fig. 1. (a) Overview of SyncSAM. (b) Multi-scale dual-branch decoder. (c) Synchronized fusion in the synchronized dual-branch encoder at the i-th stage.

scalability across different dataset sizes. (3) We extensively conduct zero-shot segmentation experiments on six external datasets, showing that SyncSAM surpasses existing medical foundation models under simple bounding box prompts, highlighting its strong potential for wide downstream applications.

2 Method

2.1 Overview

Figure. 1(a) illustrates the framework of our proposed SyncSAM, which aims to enhance medical image segmentation with a synchronized dual-branch architecture. Given a pre-processed medical image $I \in \mathbb{R}^{3\times H\times W}$ and a prompt c (points, boxes, or masks), SyncSAM predicts the corresponding binary mask $M^{(c)}$:

$$M^{(c)} = \text{SyncSAM}(I, c) \tag{1}$$

SyncSAM consists of three main components: a synchronized dual-branch encoder, a prompt encoder, and a multi-scale dual-branch decoder. The synchronized dual-branch encoder retains SAM's ViT while integrating a CNN-based synchronized fusion branch to align features at each stage [shown in Fig. 1(b)]. The prompt encoder, inherited from SAM, transforms sparse and dense prompts into vector representations, referred to as prompt token. The decoder then integrates feature maps with prompt token, while the multi-scale fusion branch [detailed in Fig. 1(c)] enhances feature refinement. Together with the proposed Med-Output Token, these components collaboratively produce the final segmentation output. The following sections detail each component.

2.2 Synchronized Dual-Branch Encoder

The synchronized dual-branch encoder serves as the image encoder in the model, encoding a preprocessed image $I \in \mathbb{R}^{3\times H\times W}$ into a rich semantic feature map $F \in \mathbb{R}^{256\times \frac{H}{16}\times \frac{W}{16}}$. To be specific, the encoder consists of two parallel branches, i.e., a ViT branch and a CNN branch. The ViT branch, inherited from SAM, captures long-range dependencies and global context while remaining frozen to limit the number of trainable parameters. In contrast, the CNN branch extracts local fine-grained features, which are particularly advantageous in handling noise and structural variations commonly found in medical images. Initially, the input image undergoes patch embedding in the ViT branch and stage 0 processing in the CNN branch as a preprocessing step. The image is then progressively processed through four consecutive stages in both branches.

To effectively integrate features from both branches, we design a synchronized fusion mechanism that merges encoded representations at each stage. At stage i, the ViT branch leverages windowed attention transformer blocks to refine the output feature map $F^{(i-1)}_{\text{output}}$ from the previous stage, producing $F^{(i)}_{\text{windowed}}$. Simultaneously, the CNN branch extracts local and domain-specific features from $F^{(i-1)}_{\text{CNN}}$, generating $F^{(i)}_{\text{CNN}}$. To ensure compatibility between the two feature maps, $F^{(i)}_{\text{CNN}}$ undergoes convolution to match the ViT feature dimensions, yielding $\tilde{F}^{(i)}_{\text{CNN}}$. This transformed CNN feature map is then point-wise added to $F^{(i)}_{\text{windowed}}$, forming the fused feature map $F^{(i)}_{\text{fused}}$. The fused representation is further refined through a global attention transformer block, producing the stage output $F^{(i)}_{\text{output}}$. Going through all four stages, the encoded feature map is further refined by the ViT neck layer, generating the final output F.

2.3 Multi-Scale Dual-Branch Decoder

The multi-scale dual-branch decoder introduces two key modifications compared to SAM's mask decoder, as shown in Fig. 1(b).

Med-Output Token. Inspired by SAM-HQ [12], we replace SAM's IoU prediction head and multiple mask tokens with a single trainable Med-Output Token (1×256). In SAM, these tokens resolve ambiguities in point-based prompts for natural images. However, such ambiguities are rare in medical image segmentation, making multiple masks and IoU prediction unnecessary.

Multi-Scale Fusion Branch. A multi-scale fusion branch is incorporated into the mask decoder, as early-stage features preserve more fine-grained edge details [22]. This branch fuses feature maps F_i from the first three ViT stages with a dimensionally adjusted F from the image encoder, generating a multi-scale feature map $\tilde{F} \in \mathbb{R}^{32\times \frac{H}{4}\times \frac{W}{4}}$. The feature fusion is performed via a series of convolutional layers. $\tilde{F}$ is then point-wise added to the mask feature F_M, which is derived from F using a Two-Way Transformer to integrate information from the output tokens, including prompt tokens and the Med-Output Token.

Table 1. The modality distributions in SA-Med2D-20M and IMed-361M.

Modality	CT	Endoscopy	PET	Fundus
Images in SA-Med2D-20M	1,645,894	4,290	5,410	1,445
Masks in SA-Med2D-20M	5,533,808	15,469	6,282	1,741
Images in IMed-361M	1,726,089	52,568	0	1,275
Masks in IMed-361M	61,190,317	52,568	0	2,348
Modality	**MR**	**Dermoscopy**	**X-ray**	**Ultrasound**
Images in SA-Med2D-20M	1,967,254	6,698	5,581	2,590
Masks in SA-Med2D-20M	5,533,808	6,858	7,067	2,590
Images in IMed-361M	410,261	6,123	566	45,103
Masks in IMed-361M	1,291,195	6,123	566	45,103

3 Experiments

3.1 Training of SyncSAM

Datasets. We train our model on two of the largest 2D medical image segmentation datasets to date: SA-Med2D-20M [29] and IMed-361M [5]. This resulted in two versions of SyncSAM, SyncSAM-SAMed and SyncSAM-IMed, to demonstrate the model's scalability across different datasets. Specifically, SA-Med2D-20M contains 3.6 million images and 15.8 million ground truth masks, while IMed-361M includes 2.2 million images and 62.6 million ground truth masks, with modality distributions as shown in the Tab. 1. For SA-Med2D-20M, which does not provide a predefined training-test split, we randomly split the data with 80% used for training and the remaining 20% for testing. For IMed-361M, we follow the official dataset split.

Loss Function. We modify SAM's loss function by removing the IoU prediction loss, as the IoU head is no longer used. The final loss function is defined as:

$$\mathcal{L} = \lambda \mathcal{L}_{Dice} + \mathcal{L}_{Focal} \tag{2}$$

Training Details. Following previous studies [13,18,6], we train SyncSAM using interactive segmentation simulation, and randomly select five masks per image for a training step. We empirically set $\lambda = 20$. For preprocessing, images are padded to square shapes and resized to 256×256 before being fed into the model. We use the ViT-B version of SAM, keeping its image encoder frozen while integrating different CNN branches—ResNet50 and ResNet34 [11]. The corresponding models, SyncSAM-50 and SyncSAM-34, contain approximately 138 million and 116 million parameters, respectively, of which about 48 million and 26 million are trainable. Training is conducted for 12 epochs with an initial learning rate of 0.0001. We use 8 NVIDIA Tesla A100 GPUs, processing 50 images and 250 masks per GPU per step.

3.2 Experimental Setup

We assess SyncSAM through three key experiments: (1) test set evaluations against models trained on the same dataset (Sec. 3.3), (2) zero-shot segmentation

Table 2. Performance comparison on the test set of SA-Med2D-20M.

Model	DSC		
	Bbox	1 pt	5 pts
SAM	66.6	24.5	52.8
MedSAM	80.8	✗	✗
SAM-Med2D	78.2	68.3	76.7
FT-SAM	74.6	61.1	73.2
SyncSAM-SAMed-34	87.0	74.1	87.6
SyncSAM-SAMed-50	**88.2**	**74.8**	**88.5**

Table 3. Performance comparison on the test set of IMed-361M.

Model	DSC		
	Bbox	1 pt	5 pts
SAM	67.2	23.3	51.4
MedSAM	83.8	✗	✗
SAM-Med2D	82.3	78.7	83.6
FT-SAM	80.1	78.3	81.9
SyncSAM-IMed-34	88.2	86.9	89.1
SyncSAM-IMed-50	**89.6**	**87.3**	**89.7**

performance on external datasets (Sec. 3.4), and (3) ablation studies analyzing the impact of individual components (Sec. 3.5). In the following tables, "SAM" denotes direct inference with SAM, while "FT-SAM" refers to SAM with only the mask decoder trained. All results are reported as Dice Similarity Coefficient (DSC) percentages, with **best** and second best performances highlighted. All reported scores are averaged over three independent runs.

3.3 Test Set Performance Comparison

In the test set comparisons, to ensure fairness, all models except SAM [13] (which performs direct inference) are trained on the same dataset. The models—MedSAM [18], SAM-Med2D [6], and FT-SAM—correspond to fully fine-tuned SAM, adapter-tuned SAM, and SAM with only the mask decoder trained, respectively. For interaction modes, we tested three types of prompts: bounding box prompt, a random point prompt, and interactive sequential point inputs (5 points). However, since MedSAM does not implement point-based interaction, it was not evaluated for this mode. The results in Tab. 2 and Tab. 3 correspond to two different test sets, demonstrating that our two versions of SyncSAM outperform all other methods across both training datasets. For MedSAM, SAM-Med2D, and FT-SAM, as the number of trainable parameters decreases, the model performance also drops to varying degrees. However, compared to the fully fine-tuned MedSAM, our model achieves higher DSC scores with less than half the number of trainable parameters, thanks to SyncSAM's model design.

3.4 Zero-Shot Segmentation Performance Comparison

We conduct zero-shot segmentation experiments on six unseen datasets: STS [25], HNTSMRG [24], CURVAS [21], COSAS24 [16], KPIs [23], and EBHI [19], which have not been used in the training sets of any of the compared models, covering five modalities in total. The compared models, besides SAM [13], include recent SOTA foundation medical image segmentation models: SAM-Med2D [6], MedSAM [18], the ScribblePrompt [26] series, and IMIS-Net [5]. To directly evaluate the zero-shot segmentation performance of each base model, we use the widely adopted bounding box prompt and perform direct inference with each model. The results are shown in Tab. 4.

Table 4. Zero-shot segmentation performance on 6 unseen datasets.

Model	MR		CT	X-ray	Microscopy	Pathology	
	HNTSMRG		CURVAS	STS	COSAS24	KPIs	EBHI
SAM	70.2		82.1	63.4	**66.2**	84.8	**80.6**
SAM-Med2D	83.7		91.5	70.1	61.9	82.8	71.6
MedSAM	62.3		65.2	59.0	59.7	70.4	77.0
ScribblePrompt-UNet	61.2		61.0	48.1	45.1	49.1	46.7
ScribblePrompt-SAM	54.2		60.2	49.4	43.4	50.6	48.3
IMIS-Net	80.6		81.7	64.6	51.6	77.4	70.5
SyncSAM-SAMed-34	83.9		92.4	**73.2**	65.9	85.8	80.0
SyncSAM-SAMed-50	84.8		93.0	72.9	64.5	**86.7**	79.2
SyncSAM-IMed-34	85.6		94.2	68.3	64.8	84.9	78.1
SyncSAM-IMed-50	**87.2**		**94.7**	68.5	62.7	85.9	78.9

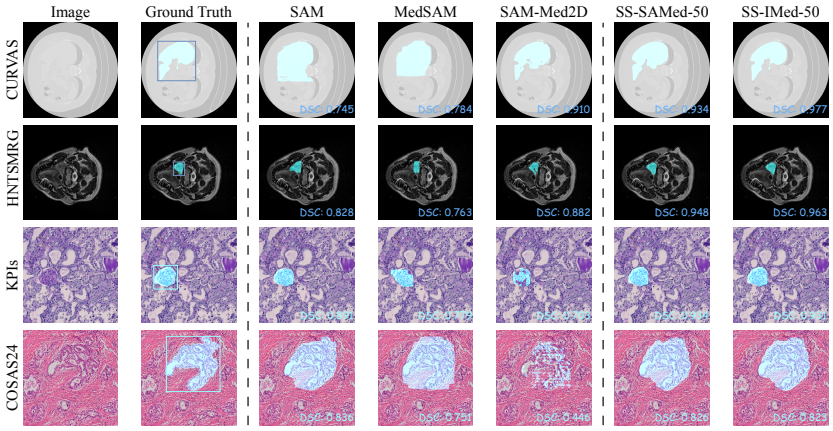

Fig. 2. Example predictions of zero-shot segmentation. SS = SyncSAM.

For MedSAM, due to its limited dataset size, its generalization to medical images remains flawed. As for the ScribblePrompt series, which focuses on training with various interaction methods, its performance under bounding box prompts is unstable. Next, we discuss the experimental results by dataset. In common large-scale datasets, CT and MR modalities typically make up about 90% of the dataset, while other modalities like X-ray, Endoscopy, Ultrasound, et al. usually account for 0.1% to 10%. Microscopy and Pathology data are usually below 0.1%, and these two modalities are even absent from the released SA-Med2D-20M and IMed-361M datasets. Therefore, we will present the results in three parts based on modality distribution. (1) For MR and CT data, the large volume of training samples enables most models to generalize well, with SyncSAM models surpassing all other comparisons. Among the SyncSAM variants, the model trained on

Table 5. Ablation. △ = single-step fusion; ★ = synchronized fusion; RN = ResNet.

#	Model	Decoder trained with		Image Encoder				DSC
		Multi-scale	Med-token	ViT (fixed)	RN-34	RN-50	Fusion	Test set
1	**SAM (fixed)**			✓				66.6
2	**FT-SAM**			✓				74.6
3	**SAM**							80.8
4				✓		✓	★	85.5
5						✓		79.1
6		✓	✓	✓				78.2
7		✓				✓		82.7
8	**SyncSAM**	✓	✓			✓		84.6
9			✓	✓		✓	★	86.6
10		✓	✓	✓	✓		★	87.0
11		✓	✓	✓		✓	△	84.9
12		✓	✓	✓		✓	★	**88.2**

IMed-361M performs better due to its larger amount of modality-specific training data. (2) For the STS dataset, where X-ray is underrepresented, most models exhibit suboptimal performance. However, SyncSAM-SAMed-34 still outperforms SAM by nearly 10%. In contrast, SyncSAM trained on IMed-361M, which contains very few X-ray samples, experiences a notable performance drop. (3) For Microscopy and Pathology data, where training images are extremely scarce or entirely absent, most models perform worse than SAM due to the reduced generalization ability after fine-tuning on medical datasets. However, SyncSAM benefits from its fully frozen ViT branch, which helps preserve SAM's encoding ability for diverse images. As a result, SyncSAM maintains a smaller performance gap compared to SAM on these datasets and in some cases even surpasses SAM.

3.5 Ablation study

We conduct ablation experiments on SA-Med2D-20M [29], using bounding boxes as prompts to evaluate the impact of each module in SyncSAM, as shown in Tab. 5. Overall, fully fine-tuned SAM achieves a DSC of 80.8 (row 3), while our synchronized encoding variant reaches 85.5 (row 4), highlighting the advantage of our architectural design. Specifically, rows 1 and 4 demonstrate that incorporating the synchronized dual-branch encoder alone improves DSC by 18.9%. Removing the ViT branch (rows 5, 7, and 8) leads to inferior performance compared to the dual-branch configuration, confirming the necessity of retaining ViT for contextual representation. Rows 11 and 12 compare different fusion strategies, showing that synchronized fusion improves DSC by 3.3%, validating its effectiveness. Furthermore, we assess the impact of decoder modifications across various settings (rows 4 vs. 9 vs. 12, rows 5 vs. 7 vs. 8, and rows 2 vs. 6), all of which demonstrate consistent performance gains from decoder enhancements. Lastly, comparing row 10 and row 12 shows that replacing ResNet-34 with ResNet-50 improves DSC by 1.2%, suggesting that our model benefits from increased capacity and is scalable with larger CNN backbones.

4 Conclusion

We introduce SyncSAM, a novel foundation model that combines the SAM with a synchronized dual-branch encoder and a multi-scale dual-branch decoder tailored for medical image segmentation. Trained on the two largest datasets, SA-Med2D-20M and IMed-361M, SyncSAM achieves SOTA performance and demonstrates strong zero-shot generalization on unseen datasets, making it a powerful baseline for medical image segmentation. In the future, we plan to probe into multi-modality foundation models for medical images.

Acknowledgments. This work was supported by NSFC under Grant 12326615, 42201394 and 12426313, and Key R&D Program of Shaanxi Province under Grant 2025CY-YBXM-040.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. An, R., Yang, S., Lu, M., Zhang, R., Zeng, K., Luo, Y., Cao, J., Liang, H., Chen, Y., She, Q., et al.: Mc-llava: Multi-concept personalized vision-language model. arXiv preprint arXiv:2411.11706 (2024)
2. An, R., Yang, S., Zhang, R., Shen, Z., Lu, M., Dai, G., Liang, H., Guo, Z., Yan, S., Luo, Y., et al.: Unictokens: Boosting personalized understanding and generation via unified concept tokens. arXiv preprint arXiv:2505.14671 (2025)
3. An, R., Zeng, K., Lu, M., Yang, S., Zhang, R., Ji, H., Zhang, Q., Luo, Y., Liang, H., Zhang, W.: Concept-as-tree: Synthetic data is all you need for vlm personalization. arXiv preprint arXiv:2503.12999 (2025)
4. Bi, H., Gao, Z., Liu, K., Song, Q., Wang, X.: Boosting few-shot remote sensing image scene classification with language-guided multimodal prompt tuning. In: 2023 International Conference on New Trends in Computational Intelligence (NTCI). vol. 1, pp. 293–297. IEEE (2023)
5. Cheng, J., Fu, B., Ye, J., Wang, G., Li, T., Wang, H., Li, R., Yao, H., Cheng, J., Li, J., et al.: Interactive medical image segmentation: A benchmark dataset and baseline. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20841–20851 (2025)
6. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sam-med2d. arXiv preprint arXiv:2308.16184 (2023)
7. Cheng, Z., Wei, Q., Zhu, H., Wang, Y., Qu, L., Shao, W., Zhou, Y.: Unleashing the potential of sam for medical adaptation via hierarchical decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3511–3522 (2024)
8. 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)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

10. Gao, Y., Xia, W., Hu, D., Wang, W., Gao, X.: Desam: Decoupled segment anything model for generalizable medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 509–519. Springer (2024)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Ke, L., Ye, M., Danelljan, M., Tai, Y.W., Tang, C.K., Yu, F., et al.: Segment anything in high quality. Advances in Neural Information Processing Systems **36**, 29914–29934 (2023)
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
14. Koleilat, T., Asgariandehkordi, H., Rivaz, H., Xiao, Y.: Medclip-sam: Bridging text and image towards universal medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 643–653. Springer (2024)
15. Lin, W., Wei, X., An, R., Ren, T., Chen, T., Zhang, R., Guo, Z., Zhang, W., Zhang, L., Li, H.: Perceive anything: Recognize, explain, caption, and segment anything in images and videos. arXiv preprint arXiv:2506.05302 (2025)
16. Liu, R., Xiao, H., Wang, Y., Zhang, M., Meng, B., Long, X., Liu, J.: Exploring domain generalization in semantic segmentation for digital histopathology: A comparative evaluation of deep learning models. In: International Conference on Biomedical Signal and Image Processing (2024)
17. Luo, Y., An, R., Zou, B., Tang, Y., Liu, J., Zhang, S.: Llm as dataset analyst: Subpopulation structure discovery with large language model. In: European Conference on Computer Vision. pp. 235–252. Springer (2024)
18. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
19. MIaMIA: EBHI-SEG (11 2022). https://doi.org/10.6084/m9.figshare.21540159.v1, https://figshare.com/articles/dataset/EBHI-SEG/21540159
20. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024)
21. Riera-Marín, M., Kleiß, J.M., Aubanell, A., Antolín, A.: Curvas dataset (Jul 2024). https://doi.org/10.5281/zenodo.12687192, https://doi.org/10.5281/zenodo.12687192
22. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
23. Tang, Y., He, Y., Nath, V., Guo, P., Deng, R., Yao, T., Liu, Q., Cui, C., Yin, M., Xu, Z., et al.: Holohisto: End-to-end gigapixel wsi segmentation with 4k resolution sequential tokenization. arXiv preprint arXiv:2407.03307 (2024)
24. Wahid, K., Dede, C., Naser, M., Fuller, C.: Training dataset for hntsmrg 2024 challenge (Jun 2024). https://doi.org/10.5281/zenodo.11199559, https://doi.org/10.5281/zenodo.11199559
25. Wang, Y., Zhang, Y., Chen, X., Wang, S., Qian, D., Ye, F., Xu, F., Zhang, H., Zhang, Q., Wu, C., et al.: Sts miccai 2023 challenge: Grand challenge on 2d and 3d semi-supervised tooth segmentation. arXiv preprint arXiv:2407.13246 (2024)

26. Wong, H.E., Rakic, M., Guttag, J., Dalca, A.V.: Scribbleprompt: fast and flexible interactive segmentation for any biomedical image. In: European Conference on Computer Vision. pp. 207–229. Springer (2024)
27. Wu, J., Wang, Z., Hong, M., Ji, W., Fu, H., Xu, Y., Xu, M., Jin, Y.: Medical sam adapter: Adapting segment anything model for medical image segmentation. Medical image analysis **102**, 103547 (2025)
28. Wu, J., Xu, M.: One-prompt to segment all medical images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11302–11312 (2024)
29. Ye, J., Cheng, J., Chen, J., Deng, Z., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sa-med2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks. arXiv preprint arXiv:2311.11969 (2023)
30. Zhu, J., Hamdi, A., Qi, Y., Jin, Y., Wu, J.: Medical sam 2: Segment medical images as video via segment anything model 2. arXiv preprint arXiv:2408.00874 (2024)