

# Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation

## 摘要

Sam模型在图像分割领域内具有卓越的性能和优秀的交互，已经成为一个备受关注的模型。许多从业人员认为，Sam已经“完成”了图像分割任务。最近的研究表明Sam在医学图像分割方面的表现不尽如人意，Sam无法有效应用在医学图像分割领域最主要的一个原因是，缺乏足够的医疗训练数据。为了扩展 Sam 到医学图像分割任务中，本文提出了 Medical SAM Adapter (MSA)，将医学特定领域知识通过采用了一种名为 Adaption 的参数高效微调技术集成到分割模型中，而不是仅对 Sam 模型微调。MSA 在 19 个医学图像分割任务中展现了卓越的表现，包括 CT、MRI、超声图像、眼底图像和皮肤镜像。与 nnUNet、TransUNet、UNetr、MedSegDiff 等各种先进的医学图像分割方法相比，MSA 表现出色，甚至超越了完全微调的 MedSAM。

**关键词：**医学图像分割；自适应技术；高效微调

## 1 引言

最近的研究对分割一切模型（Sam）[\[10\]](#)在医学图像分割任务中的应用进行了扩展。尽管Sam在自然图像分割方面表现出色，但在医学图像上的性能却不尽人意。本文提出了一种最小化参数的微调技术——适配（Adaption）[\[8\]](#)，通过在预训练的Sam模型中插入适配器模块以实现参数高效的微调，这样只需调整少量参数即可[\[9\]](#)。作者认为这种方法适合将Sam应用于医学图像，因为它不仅能避免灾难性遗忘，还能在数据稀缺情况下更好地泛化。研究成果表明，经过适配的Sam模型（MSA）在19项医学图像分割任务上的表现优于现有最先进的方法，证明了这种技术路线的有效性。这项工作推动了Sam向“分割任何事物”的最终目标迈进了一大步，尤其是在针对医学图像分割的应用方面取得了显著进展。

本文主要贡献包括：

(1) 将流行且强大的 Sam 模型扩展到医学领域，这是朝着“分割任何东西”的终极目标迈出的重要一步。

(2) 首次提出将适应性方法用于医学图像分割。作者在设计适配器时考虑了领域特定的知识，例如医学数据的高维度(3D)和解码器的点击和目标框提示的独特设置。

(3) 在 19 个医学图像分割任务中评估了 MSA 模型，包括 MRI、CT、眼底图像、超声图像和皮肤镜图像等不同的图像模态。结果表明，MSA 在性能上显著优于以前的最先进方法。

## 2 相关工作

### 1. 为什么需要 Sam 模型来进行医学图像分割任务?

交互式分割是所有分割任务的一种范式, prompt决定了预期结果的细粒度, 对于zero-shot任务而言是必须的 [5], 并且需要由用户提供。例如, 在医学图像任务上, 根据不同的要求和用途, 需要从一张眼底图像中分割出不同目标 [7], 比如血管、视盘、视杯和黄斑。Sam为交互式分割提供了一个很好的框架, 使其成为实现基于prompt的医学图像分割的完美起点 [1]。

### 2. 为什么需要对模型进行微调?

Sam 已经在最大的分割数据集上训练过, 预训练模型是有价值的, 因为许多研究表明, 在自然图像上预训练也对医学图像分割有益, 至少在收敛速度上 [11]。

### 3. 为什么选择PEFT和Adaption?

PEFT已经被证明是一种有效的策略 [12], 可以针对特定用途对大型的FM进行微调, 与完全微调相比, 它保持了大部分参数的冻结, 学习的参数量明显减少, 通常少于总数的5%, 这就使得效率更高。研究也表明了, PEFT方法比完全微调更好, 因为它能够避免灾难性遗忘, 也能够更好地推广到域外场景, 尤其是在数据量不足的情况下 [13], 在所有的PEFT中, Adaption是一个有效的工具, 不仅在NLP中, 在CV领域也很有效, Adaption可以很容易被用在各种下游的CV任务上, 所以作者认为Adaption是将Sam用到医学领域的最合适的技术 [6]。

## 3 本文方法

### 3.1 Sam架构

Sam主要包括三个部分: 一个图像编码器, 一个prompt编码器, 一个mask解码器 [10]。图像编码器是基于由MAE预训练的标准ViT, 采用的是ViT-H/16的变体, 用 $14 \times 14$ 的窗口注意力和4个等距的全局注意力块, 图像编码器的输出是输入图像的16倍下采样的嵌入向量。Prompt编码器可以是稀疏的(点、框、文本)也可以是密集的(mask), 本文只关注稀疏编码器, 将点和框表示为位置编码, 对于每个prompt类型, 将前面学习到的嵌入向量和位置编码相加。Mask解码器是一个经过修改的包括一个动态的mask预测头的Transformer解码器块 [2]。Sam使用双向交叉注意力机制, 一个用于prompt到图像的嵌入, 另一个用于图像到prompt的嵌入, 在每个块中学习prompt和图像嵌入向量之间的交互, 在运行两个区块后, Sam对图像嵌入向量上采样, 再经过一个MLP将输出的token映射到动态的线性分类器, 预测给定图像的目标mask。

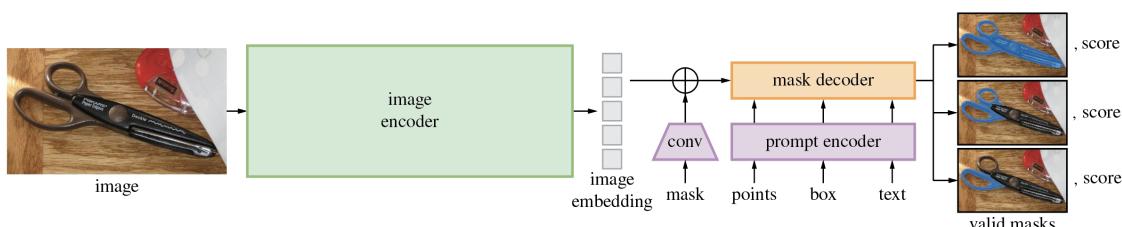


图 1. Sam网络结构

### 3.2 MSA架构

作者在 SAM 网络架构的特定位置插入了一个 Adapter 模块，来实现 fine-tune。

Adapter 是一个瓶颈模型，如图2(b)所示，组合顺序为下投影、ReLU 激活和上投影。下投影使用简单的 MLP 层将给定的嵌入压缩到较小的维度。上投影使用另一个 MLP 层将压缩的嵌入扩展回其原始维度。

在 SAM 编码器中，如图2(b)所示，每个 ViT 块部署了两个适配器，第一个适配器位于多头注意力后，第二个适配器位于 MLP 层的残差路径上。

在 SAM 解码器中，如图2(d)所示，每个 ViT 块部署了三个适配器，分别是 prompt-to-image 嵌入的多头交叉注意力后、MLP-enhanced 嵌入后和图像嵌入到 prompt 交叉注意力的残差连接后。

为了适应医学图像中的三维特点，作者提出了一种基于图像到视频适应的新颖方法，如图2(c)所示，在每个块中将注意力操作分为空间分支和深度分支。

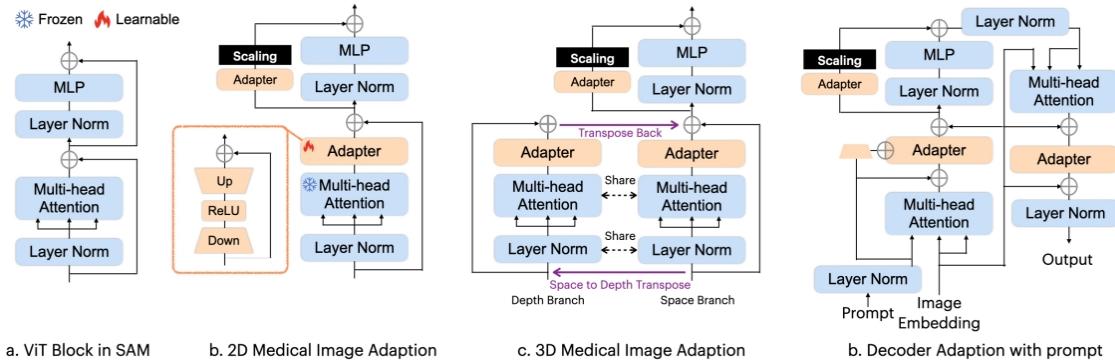


图 2. MSA网络结构

### 3.3 训练策略

#### 3.3.1 编码器的预训练

遵循SAM的思路，作者使用医学影像对编码器进行预训练，首先使用了四个医学图像数据集的混合数据，包括RadImageNet数据集，数据包括了135万张放射学图像（CT、MRI、US），涵盖了踝部/足部、脑部、髋部、膝盖、肩膀、脊柱、腹部、骨盆、胸部和甲状腺。EyePACSp数据集包括88702张彩色眼底图像，由多个初级护理诊所在不同设备环境下采集。BCN-20000和HAM-10000数据集包括了约30000张患有黑色素瘤或者黑痣的皮肤显微镜图像，所有这些数据都是公开的 [3]。与SAM中对MAE的预训练不同，作者使用了几种自监督学习方式的组合进行预训练，其中前两种方法是对比嵌入混合（e-Mix）和洗牌嵌入预测（ShED）。e-Mix是一种对比目标，对一批原始输入嵌入进行加法混合，用不同的系数加权处理，然后训练一个编码器，使其生成一个混合向量，该向量与原始输入的嵌入向量按照混合系数，成比例地靠近。ShED对一部分嵌入信息洗牌（打乱），用一个分类器训练编码器，以预测哪些嵌入信息被洗牌操作扰动了。然后还用了MAE，遵循SAM的原始实现，对给定遮掩比例的输入嵌入进行掩码处理，并训练模型去重建他们 [4]。

### 3.3.2 带Prompt的训练

这个过程与SAM基本相同，在新的数据集上微调带prompt的SAM，但是有一部分的修改。对于点击提示，正点击表示前景区域，负点击表示背景区域，使用随机和迭代的点击采样策略的组合进行训练。具体来说，首先随机采样进行初始化，然后使用迭代采用程序添加一些点击，迭代采样策略类似于与真实的用户进行交互，因为在实际运用中，每个新的点击都会被放置在网络预测的错误的区域。作者使用了与SAM不同的文本prompt训练策略 [5]，在SAM中是采用由CLIP生成的目标对象的图像嵌入作为其在CLIP中的相应文本描述或者定义密切相关的图像嵌入，然后由于CLIP几乎没有在医学图像数据集上进行训练，很难将图像上的器官/病变与相应的文本定义联系起来。本文中，作者先从ChatGPT中随机生成包含目标（即视盘、脑肿瘤）的定义作为关键词的多个自由文本，并利用CLIP提取文本嵌入作为训练的prompt，一个自由文本可以包括多个目标，这种情况下就使用相应的mask去监督模型 [6]。

## 4 复现细节

### 4.1 数据集

五个医学图像分割数据集，分为两类，一类测试整体分割性能和SOTA对比，选择腹部多器官分割，AMOS2022和BTCV；其余四个任务分别用于验证对不同形式和类型任务的推广，包括眼底图像上的视盘和视杯分割（REFUFE和RIGA）、脑 MRI 图像上的脑肿瘤分割（BraTs）、超声图像上的甲状腺结节分割（TNSCUI和DDTI）以及皮肤镜图像上的黑色素瘤或痣分割（ISIC）。

### 4.2 主要结果

对比SAM，MSA，MedSAM（在医学图像上finetune SAM）；评价指标使用dice。

Table 1: The comparison of MSA with SOTA segmentation methods and original SAM over AMOS dataset evaluated by Dice Score. Best results are denoted as **bold**.

Methods	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	Panc.	RAG	LAG	Duo.	Blad.	Postc.	Avg
TransUNet	0.881	0.928	0.919	0.813	0.740	0.973	0.832	0.919	0.841	0.713	0.638	0.565	0.685	0.748	0.692	0.792
EnsDiff	0.905	0.918	0.904	0.732	0.723	0.947	0.838	0.915	0.838	0.704	0.677	0.618	0.715	0.673	0.680	0.786
SegDiff	0.885	0.872	0.891	0.703	0.654	0.852	0.702	0.874	0.819	0.715	0.654	0.632	0.697	0.652	0.695	0.753
UNetr	0.926	0.936	0.918	0.785	0.702	0.969	0.788	0.893	0.828	0.732	0.717	0.554	0.658	0.683	0.722	0.762
Swin-UNetr	0.959	0.960	0.949	0.894	0.827	<b>0.979</b>	0.899	0.944	0.899	0.828	0.791	0.745	0.817	0.875	0.841	0.880
nnUNet	0.965	0.959	0.951	0.889	0.820	0.980	0.890	0.948	0.901	0.821	0.785	0.739	0.806	0.869	0.839	0.878
MedSegDiff	0.963	<b>0.965</b>	0.953	0.917	0.846	0.971	0.906	0.952	0.918	0.854	0.803	0.751	<b>0.819</b>	0.868	0.855	0.889
SAM 1 point	0.632	0.759	0.770	0.616	0.382	0.577	0.508	0.720	0.621	0.317	0.085	0.196	0.339	0.542	0.453	0.493
SAM 3 points	0.733	0.784	0.786	0.683	0.448	0.658	0.577	0.758	0.625	0.343	0.129	0.240	0.325	0.631	0.493	0.542
SAM 10 points	0.857	0.855	0.857	0.800	0.643	0.811	0.749	0.842	0.677	0.538	0.405	0.516	0.480	0.789	0.637	0.699
MedSAM 1 point	0.671	0.803	0.825	0.687	0.541	0.712	0.671	0.785	0.703	0.607	0.531	0.588	0.729	0.814	0.833	0.700
MSA 1-point	<b>0.968</b>	0.961	<b>0.959</b>	<b>0.926</b>	<b>0.861</b>	0.971	<b>0.919</b>	<b>0.960</b>	<b>0.928</b>	<b>0.863</b>	<b>0.825</b>	<b>0.767</b>	0.803	<b>0.879</b>	<b>0.862</b>	<b>0.893</b>

Table 2: The comparison of MedSegDiff-V2 with SOTA segmentation methods over BTCV dataset evaluated by Dice Score. Best results are denoted as **bold**.

Model	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	Veins	Panc.	AG	Ave
TransUNet	0.952	0.927	0.929	0.662	0.757	0.969	0.889	0.920	0.833	0.791	0.775	0.637	0.838
EnsDiff	0.938	0.931	0.924	0.772	0.771	0.967	0.910	0.869	0.851	0.802	0.771	0.745	0.854
SegDiff	0.954	0.932	0.926	0.738	0.763	0.953	0.927	0.846	0.833	0.796	0.782	0.723	0.847
UNetr	0.968	0.924	0.941	0.750	0.766	0.971	0.913	0.890	0.847	0.788	0.767	0.741	0.856
Swin-UNetr	0.971	0.936	0.943	0.794	0.773	0.975	0.921	0.892	0.853	0.812	0.794	0.765	0.869
nnUNet	0.942	0.894	0.910	0.704	0.723	0.948	0.824	0.877	0.782	0.720	0.680	0.616	0.802
MedSegDiff	0.973	0.930	0.955	0.812	0.815	0.973	0.924	0.907	0.868	0.825	<b>0.788</b>	0.779	0.879
SAM 1 points	0.518	0.686	0.791	0.543	0.584	0.461	0.562	0.612	0.402	0.553	0.511	0.354	0.548
SAM 3 points	0.622	0.710	0.812	0.614	0.605	0.513	0.673	0.645	0.483	0.628	0.564	0.395	0.631
SAM 10 points	0.785	0.774	0.863	0.658	0.673	0.785	0.760	0.712	0.562	0.703	0.651	0.528	0.704
MedSAM 1 point	0.751	0.814	0.885	0.766	0.821	0.901	0.855	0.872	0.746	0.771	0.760	0.705	0.803
MSA 1 point	<b>0.978</b>	<b>0.935</b>	<b>0.966</b>	<b>0.823</b>	<b>0.818</b>	<b>0.981</b>	<b>0.931</b>	<b>0.915</b>	<b>0.877</b>	<b>0.811</b>	0.767	<b>0.809</b>	<b>0.883</b>

图 3. MSA在数据集AMOS和BTCV上的表现

不论给定的提示是什么，SAM 在目标医疗图像分割任务中的零样本性能一般都不如完全训练有素的模型。虽然这种比较似乎不公平，SAM 的零样本性能已被证明优于完全训练的自然图像数据集模型。这表明 SAM 在医学图像上具有较差的零样本转移能力，这在许多其他研究中也观察到。

通过比较不同提示语的性能，可以发现三点提示语的性能略优于一点提示语，而十点提示语的性能则明显优于三点提示语。尽管10点提示仍然比其他训练有素的模型表现差，但考虑到零样本设置，这仍然优秀。这突出了 SAM 架构在医学图像方面的巨大潜力。通过利用这种潜力，MSA 比只给出1点提示的 SAM 获得了显著的改进。在 AMOS 数据集上，MSA 在15个器官中的12个上实现了 SOTA 性能，并且总体性能最好。在 BTCV 数据集上，MSA 在12个器官中的11个上实现了 SOTA 性能，并且整体性能最好。这些结果表明，通过使用正确的微调技术和一个超级好的预先训练的模型，即使是在自然图像上，也可以对医疗图像分割非常有益，甚至可以提升其性能超过专门优化的医疗图像分割模型。

	Optic-Cup		Brain-Tumor			Thyroid Nodule	
	Dice	IoU	Dice	IoU	HD95	Dice	IoU
ResUnet	80.1	72.3	78.4	71.3	18.71	78.3	70.7
BEAL	83.5	74.1	78.8	71.7	18.53	78.6	71.6
TransBTS	85.4	75.7	87.6	78.44	12.44	83.8	75.5
EnsemDiff	84.2	74.4	88.7	80.9	10.85	83.9	75.3
MTSeg	82.3	73.1	82.2	74.5	15.74	82.3	75.2
UltraUNet	83.1	73.78	84.5	76.3	14.03	84.5	76.2
SegDiff	82.5	71.9	85.7	77.0	14.31	81.9	74.8
nnUNet	84.9	75.1	88.5	80.6	11.20	84.2	76.2
TransUNet	85.6	75.9	86.6	79.0	13.74	83.5	75.1
UNetr	83.2	73.3	87.3	80.6	12.81	81.7	73.5
Swin-UNetr	84.3	74.5	88.4	81.8	11.36	83.5	74.8
MedsegDiff	85.9	76.2	88.9	81.2	10.41	84.8	76.4
SAM 1 points	-	-	63.2	58.6	25.53	-	-
SAM 3 points	-	-	65.5	61.7	24.87	-	-
MSA	<b>86.8</b>	<b>78.8</b>	87.6	81.2	12.46	<b>86.3</b>	<b>78.7</b>

图 4. MSA 与 SAM 和 SOTA 分割方法在不同图像模式下的比较

作者还比较了 MSA 和最先进的(SOTA)分割方法提出的三个具体任务与不同的图像模式。MSA 在视杯和甲状腺结节分割任务上获得了 SOTA 性能，并且在脑肿瘤分割上优于大多数模型，显示了它对各种医学分割任务和图像模式的一般化能力

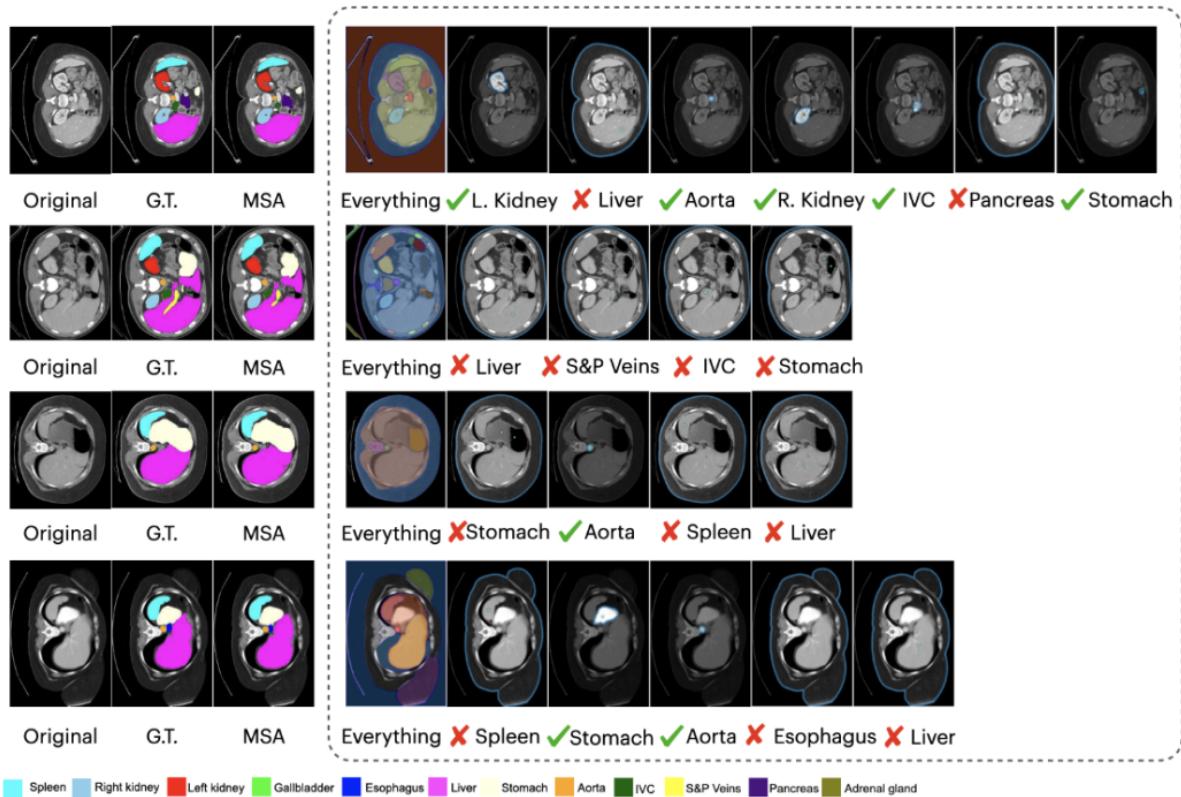


图 5. MSA 和 SAM 对腹部多器官分割的视觉比较

MSA 在难以被人眼识别的部位进行了准确的分割，而 SAM 则在一些器官其实很清晰的情况下失败了。这再次说明了，在医学图像分割中，对于一个通用分割模型进行精细调整是非常必要的。

## 5 实验结果分析

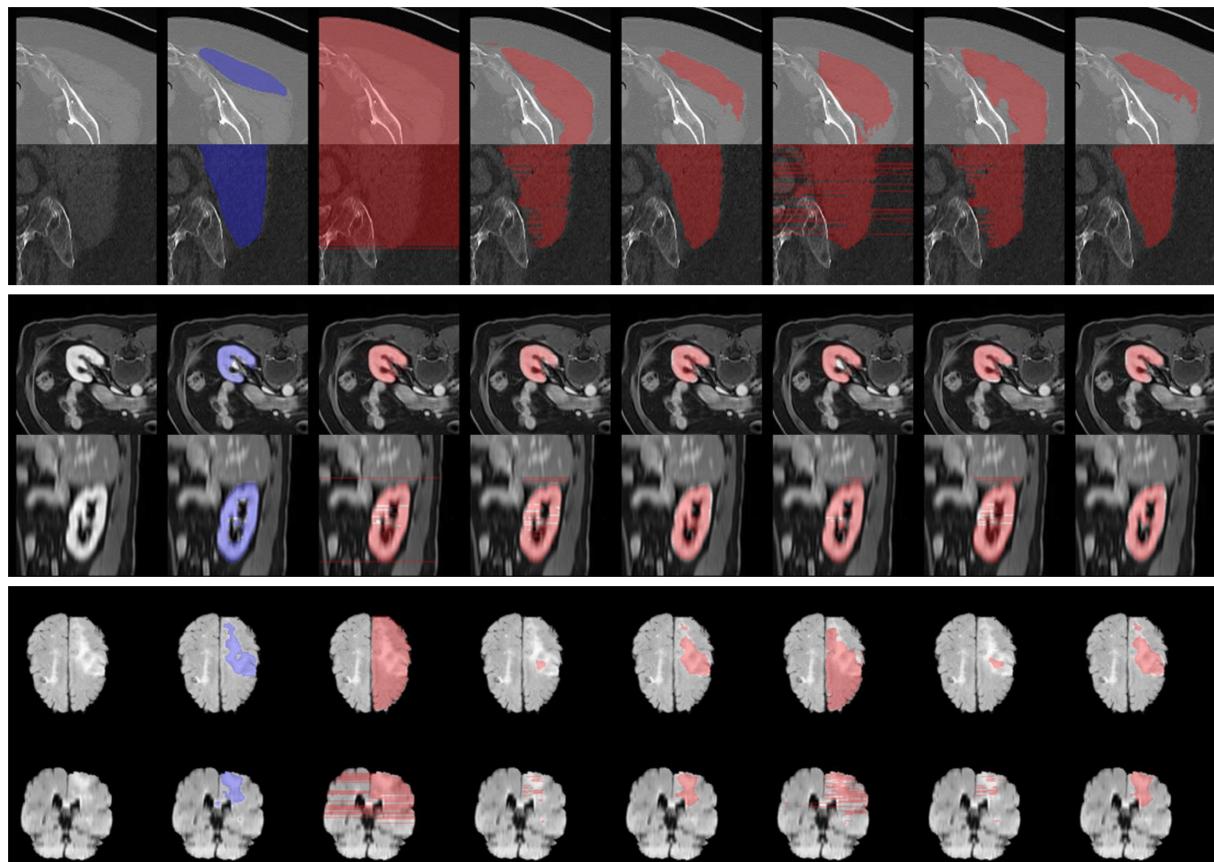


图 6. 腹部多器官分割结果

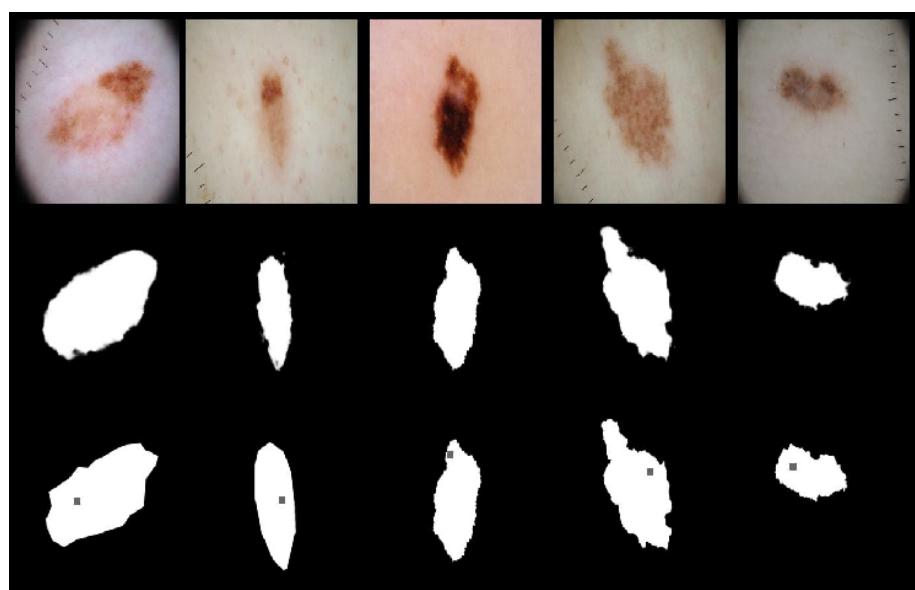


图 7. 皮肤黑色素瘤图像分割 (2D)

## 6 总结与展望

作者将通用分割模型 SAM 扩展到医学图像分割领域，并命名为 MSA。通过采用参数有效的适应性技术，一种成本效益的微调技术，在 19 个医学图像分割任务中实现了显著的改进，并在 5 种不同的图像模态下取得了 SOTA 性能。这些结果证明了我们的适应性方法对于医学图像的适应性是有效的，同时也表明了将通用的分割模型用于医学应用的潜力。本复现工作在AMOS、BTCV等数据集上测试模型分割性能,评估指标包括Dice系数、IoU等,在2d和3d医学图像分割数据集上,复现出了论文报告的指标数值,例如在AMOS数据集上单点提示下达到85.3%的平均Dice。复现后的代码实验效果能大体接近论文所示的实验结果。希望这篇工作可以成为推进通用医学图像分割的起点，并激发新的微调技术的发展。

## 参考文献

- [1] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021.
- [2] J Chen, Y Lu, and Q TransUNet Yu. Transformers make strong encoders for medical image segmentation. arxiv 2021. *arXiv preprint arXiv:2102.04306*.
- [3] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.
- [4] Ruining Deng, Can Cui, Quan Liu, Tianyuan Yao, Lucas W Remedios, Shunxing Bao, Bennett A Landman, Lee E Wheless, Lori A Coburn, Keith T Wilson, et al. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging. *arXiv preprint arXiv:2304.04155*, 2023.
- [5] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 9224–9232, 2018.
- [6] 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.
- [7] Sheng He, Rina Bao, Jingpeng Li, P Ellen Grant, and Yangming Ou. Accuracy of segment-anything model (sam) in medical image segmentation tasks. *arXiv preprint arXiv:2304.09324*, 2023.
- [8] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023.
- [9] Sabarinath Mahadevan, Paul Voigtlaender, and Bastian Leibe. Iteratively trained interactive segmentation. *arXiv preprint arXiv:1805.04398*, 2018.

- [10] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review*, page 21260, 2008, Jun 12.
- [11] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32, 2019.
- [12] Saikat Roy, Tassilo Wald, Gregor Koehler, Maximilian R Rokuss, Nico Disch, Julius Holzschuh, David Zimmerer, and Klaus H Maier-Hein. Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model. *arXiv preprint arXiv:2304.05396*, 2023.
- [13] Wenzuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I* 24, pages 109–119. Springer, 2021.