其训练数据中的对象大多有明显的边界。这并不适用于医学图像，因为肿瘤与其周围组织之间的边界通常是模糊的。因此，需要适应或重新设计模型以将 SAM 的识别能力转移到考虑领域知识的特定应用程序中。最近关于参数适应方法的工作试图通过学习特定于任务的视觉提示来更新预先训练的参数，指定要调优的少量参数，或合并轻量级的即插即用适配器。即使只微调或添加少量参数，也取得了可喜的成果。本文寻求一种有效的 3DSAM-adapter 版本模型用于医学成像。

2 相关工作

2.1 计算机视觉中的基础模型

随着深度学习模型的突破，大多数现代视觉模型都遵循预训练和微调范式。大型和可推广的基础模型在计算机视觉中引起了极大的兴趣，这种方法受益于包括自监督学习、对比学习、语言视觉预训练等多种预训练技术。最近，在超过 11M 图像上预训练的 SAM 作为图像分割的通用基础大模型，并以交互和可提示的方式在野外分割任何东西具有强大的零镜头能力。并发工作之一 SEEM 提出了一种更通用的提示方案来支持语义感知开放集分割。SegGPT 进一步追求图像或视频中的板上下文分割任务。

2.2 参数高效的模型微调

作为基础模型的广泛使用，参数高效微调的主题引起了很多关注。现有的高效调优方法可以分为三类。其一是在原始模型中插入轻量级适配器或提示，并且只调整这些参数。其二是基于规范的方法，选择小部分原始参数进行调整。其三是基于重新参数化的方法，即使用低秩矩阵来近似参数更新。最近，一些工作将预先训练好的图像模型应用于视频理解或体积分割。然而，这些方法将附加维度解释为“词组”，并使用特殊模块来聚合该维度的信息。本文认为需要考虑所有三个维度都是各向同性的，并且选择直接调整训练好的 transformer 块以捕获 3D 信息。

2.3 医学成像中的肿瘤分割

肿瘤分割是计算机辅助医学图像分析中最常见但最具挑战性的任务之一。深度神经网络的最新成果对肝脏、肾脏胰腺和结肠等不同解剖区域应用的性能改进做出了重大贡献。然而，即使对于目前最先进的分割网络，如 nnU-Net、UNETER++和 3D UX-Net，精确的肿瘤分割仍然具有挑战性，因为肿瘤通常具有小尺寸、不规则形状、低对比度和模糊边界的显著特性。不出所料，在最近报道的医学图像 SAM 应用中，与其他解剖结构(如 3D 器官)相比，SAM 在肿瘤分割任务上获得了更差且不稳定的结果。针对这一问题本文提出的用于肿瘤分割场景

的适配器，以解决原始 SAM 最显着的弱点。

3 本文方法

3.1 本文方法概述

本文提出了一种新的参数高效自适应方法——3DSAM-adapter，将 SAM 整体从 2D 调整为3D进行医学图像分割。首先，对于处理输入数据的图像编码器精确地设计了修改方案，允许原始的 2D transformer 支持 3D 体积输入，同时保持尽可能多的预先训练的权重以便重用。另外，本文发展在 2D 图像上预训练的权重仍然可以通过参数高效的微调来捕获一些 3D 空间模式。其次，对于提示编码器部分，本文建议使用图像嵌入的视觉采样器作为点提示表示，而不是使用位置编码作为提示表示，并进一步使用一组全局查询来消除噪声提示。该策略被证明可以很好地克服维度提升时图像标记大小急剧增加引起的过度平滑问题，且有效提高了模型对不准确提示的鲁棒性。最后，对于输出部分的掩码解码器，本文模型采用轻量级设计，添加多层聚合。本文中对医学肿瘤分割数据集进行了实验，与包括 nn-UNet 在内的领域 SOTA 方法以及最近的适配器进行了综合比较。结果表明，本文的方法可以大大优于现有方法。该方法还显示了对提示的数量和位置的鲁棒性。3DSAM-adapter 模型方法的主要贡献总结如下：

精心设计一种基于原始 SAM 架构的修改方案，提出了一种支持模型整体从 2D 到 3D 的适应方法，它只增加了 7.79%的参数，并保持大部分预先训练的权重可重复使用，同时在体积医学图像分割中表现良好。

引入了一种新的参数高效微调方法，以有效地利用在 2D 图像上预训练的大型图像模型进行 3D 医学图像分割，原始模型只有 16.96% 的可调参数（包括新添加的参数）。

对四个数据集进行了实验，用于医学图像分割。结果表明， 3DSAM-adapter 在四个数据集中的三个（肾肿瘤为 8.25%，胰腺肿瘤为 29.87%，结肠癌为 101.11%）上显着优于 nn-UNet，在肝脏肿瘤上具有可比性。另外还展示了该方法相对于最近的参数高效微调方法的卓越性能。

3.2 3DSAM-adapter 模型架构

本文基于两个标准的自适应方法考虑：

1）使模型能够直接学习 3D 空间模式；

2）从预训练模型继承大部分参数，以及扩大小尺寸的增量参数且易于微调。

模型的细节如图 1 所示：

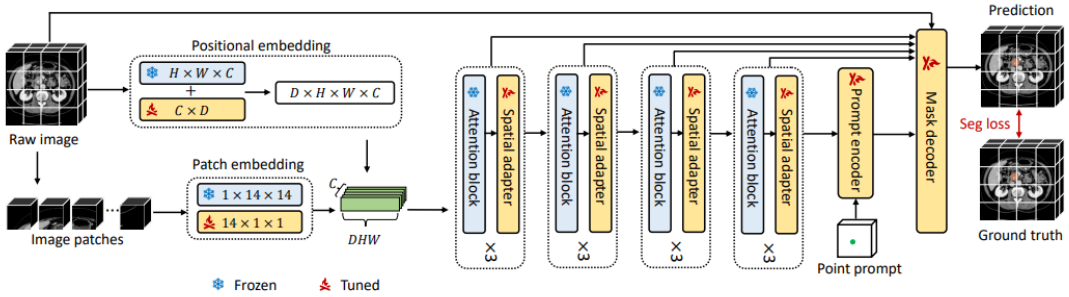

Figure 1: Overview of our proposed method for 3DSAM-adapter. The original ViT is modified to support volumetric inputs. The prompt encoder is redesigned to support 3D point prompt, and the mask decoder is updated to 3D CNN with multi-layer aggregation to generate 3D segmentation.

图 1 3DSAM-adapter 模型架构

原始的 SAM 是基于变压器，它包含多个 attention 块，从而支持不同 tofen 大小的输入。同时，体积医学图像通常是各向同性的，像素之间的空间关系与 2D 情况非常相似。因此，我们假设经过训练以学习 2D 空间特征的网络也可以轻松适应捕获 3D 模式。剩下的唯一事情是如何初始化 3D 补丁的标记以及如何以参数有效的方式通知新位置信息的模型。具体来说，网络的每个模块修改如下：

Patch embedding。利用 1 × 14 × 14 和 14 × 1 × 13D 卷积的组合作为 14 × 14 × 14 卷积的近似。用预先训练的 2D 卷积的权值初始化 1 × 14 × 14 卷积，并在微调阶段保持冻结。对于新引入的 14 × 1 × 1 3D 卷积，深度卷积用于进一步减少可调参数的数量。

Positional encoding。预训练的 ViT 包含大小为 c × H × W 的查找表和位置编码。我们还用零填充大小为 c × D 的可调查找表。3D 点(d, h, w)的位置编码可以是冻结查找表中嵌入的总和，其中 (h, w)和嵌入 (d)的可调查找表中。

Attention block。可以直接修改注意块以适应 3D 特征。对于 2D 输入，查询的大小为 [B, HW, c]，它可以很容易地适应为 [B, DHW, c]，用于继承所有预训练权重的 3D 输入。另外本文使用与 Swin Transformer 类似的滑动窗口机制来减少维度提升引起的内存成本。

Bottleneck。由于卷积层通常比 Transformer 更容易优化，本文将 Bottleneck 中的所有 2D 卷积替换为 3D 卷积，并从头开始训练它们。

通过上述修改即可优雅地将 2D ViT 升级为 3D ViT，同时保持大部分参数可重用。完全微调 3D ViT 可能是内存密集型的。为了解决这个问题，本文建议利用轻量级适配器进行有效的微调。第一个提出的适配器由下投影线性层和上投影线性层组成，该架构如图 2 所示：

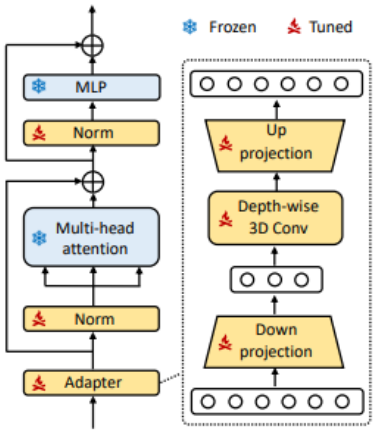

Figure 2: Spatial adapter.

图 2 Spatial adapter 架构

3.3 基于视觉采样的提示编码器

原始 SAM 利用位置嵌入来表示提示。基于傅立叶特征的位置嵌入方式应用于提示和图像，使相同位置对应的提示和图像嵌入具有相同的位置编码。然后，提示嵌入与图像嵌入交叉注意，从而将纯位置特征转换为语义特征。这种交叉注意力适用于 2D 情况，但在应用于 3D 特征图时可能会导致过度平滑问题。提升到 3D 会导致令牌数的灾难性上升，因此概率趋于均匀分布，这使得提示嵌入难以充分提取语义信息。

为此，本文使用视觉采样器而不是位置编码来表示提示。整个过程如图 3 所示。

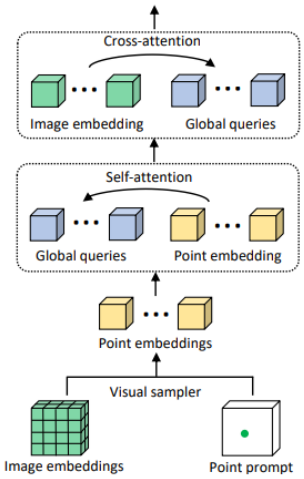

Figure 3: Structure of our prompt encoder based on visual sampler and global queries cross-attention.

图 3 提示编码器架构

给定点的坐标，本方法从特征图直接插值以获取嵌入，从而保证提示与图像嵌入共享相

同的语义特征。该过程将几个标记初始化为全局查询，然后在全局查询和这些点嵌入之间应用自注意力。之后，仅将图像嵌入的交叉注意力应用于这些全局查询。由于点提示和全局查询的数量都非常小，这可以缓解过度平滑的问题。此外，这可能会对噪声点带来更高的容忍度，因为全局查询可以作为原型，只有具有特定特征的点嵌入将具有很高的相似性。

3.4 轻量级的掩码解码器

原始 SAM 的掩码解码器被设计为轻量级的，包含一堆卷积层。本文方法用 3D 卷积替换所有 2D 卷积，以直接生成 3D 掩码。初始解码器在没有任何渐进上采样或跳过连接的情况下设计。这对于物体大小通常很大且边界清晰的自然图像效果很好。然而，对于体积医学图像分割，人们普遍认为在多个层次上具有跳跃连接的 U 形网络至关重要，因为医学图像中的对象通常很小并且具有模棱两可的边界，这需要具有更高分辨率的细节来更好地区分。

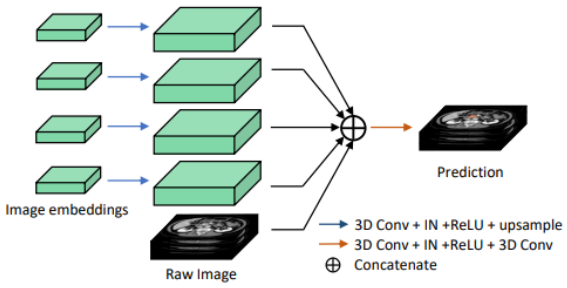

Figure 4: Structure of our lightweight mask decoder with multi-layer aggregation.

图 4 掩码解码器架构

为了缓解上述问题同时保持轻量级属性，本文在解码器中使用了多层聚合机制，其中编码器的中间输出连接在一起以产生掩码特征图，整个结构始终保持轻量级。为了更好地利用来自原始分辨率的信息，在将掩码特征图上采样到原始分辨率后，将其与原始图像连接起来，并使用另一个 3D 卷积来融合信息并生成最终掩码。该部分结构如图 4 所示。

总体而言，本文引入了一种整体方案来适应 SAM 进行医学图像分割。图像编码器的适应侧重于如何有效地利用预训练的参数来提取 3D 空间特征，而提示编码器和掩码解码器的修改主要解决维度提升带来的计算问题，也有助于模型更好地与特定领域的需求保持一致。将这两种变换 SAM 组合成一个潜在的强大的体积医学图像分割工具。

4 复现细节

4.1 与已有开源代码对比

本次复现中，参考了作者在 github 网站 https://github.com/med-air/3DSAM-adapter 所提供的代码，使用 python 实现了基于单个或多个点提示模式的 3DSAM-adapter 模型，完成了

针对胰腺肿瘤的精准分割。

4.2 数据集

本文所使用的数据集为 MSD Pancreas Tumour 数据集，该数据集是 MSD (Medical Segmentation Decathlon, 医学图像分割十项全能挑战赛) 中的 Task07 即第 7 个子任务，目标是从 CT 图像中分割出胰腺和肿瘤，MSD 选择该数据集的原因是“标签不平衡，其中包括大（背景）、中（胰腺）和小（肿瘤）结构”。胰腺肿瘤分割也是 MSD 最难的两项任务之一（另一个是结肠肿瘤分割）。该数据集包含 420 例 3D CT 数据，官方划分为 281 例训练集和 139 例测试集，其中测试集可以通过官网提交分割结果进行测试。该数据集包括三种类型的胰腺肿瘤——管腺内粘液瘤、胰腺神经内分泌瘤和胰腺导管腺瘤。

4.3 性能指标

本实验用到的性能指标为 DSC (Dice Similarity Coefficient)和 Dice 系数，二者都是用来评估两个集合的相似性的指标。

DSC 是一种衡量两个集合重叠程度的指标，通常用于评估图像分割任务的性能。它的计算公式如下：

$$Dice = \frac{2TP}{FP + 2TP + FN}$$

5 实验结果分析

本次复现实验使用 MSD Pancreas Tumour 数据集进行训练，模型最终测试结果如表一所示。从表 1 结果可以看出，在单位体积内给定不同数量的点作为提示，也对模型的分割效果有一定影响，且点的数量越多，分割效果越好。

表 1 在 MSD 数据集上的复现结果

Pancreas Tumor

Methods	Dice	NSD
3DSAM-adapter(1 pt/volume)	42.43	65.20
3DSAM-adapter(3 pt/volume)	46.17	67.43
3DSAM-adapter(10 pt/volume)	49.56	71.82

实验结果可视化如下：

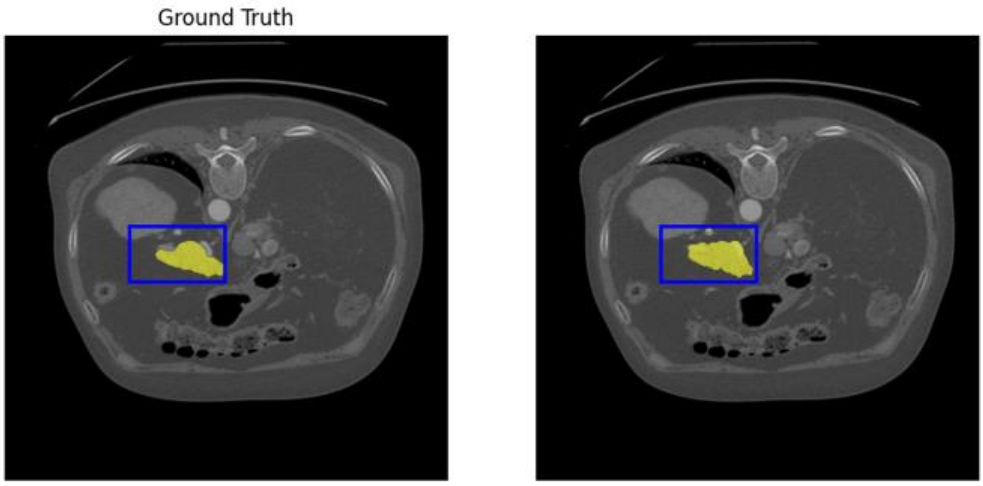

图 5 可视化结果

图 6 为原论文中的实验结果，包含 nnU-Net、3D UX-Net 等先进方法以及本文方法在四个数据集上的表现。红框部分可以看出，3DSAM-adapter 模型在胰腺肿瘤分割任务中达到了 SOTA。另外，本次复现实验的结果与论文所展示的结果相比相对接近，也取得了良好的分割性能。

Methods	Kidney Tumor		Pancreas Tumor		Liver Tumor		Colon Cancer		#Tuned Params
	Dice ↑	NSD ↑	Dice ↑	NSD ↑	Dice ↑	NSD ↑	Dice ↑	NSD ↑	
nnU-Net (Nat. Methods 2021) [22]	73.07	77.47	41.65	62.54	**60.10**	**75.41**	43.91	52.52	30.76M
TransBTS (MICCAI 2021) [52]	40.79	37.74	31.90	41.62	34.69	49.47	17.05	21.63	32.33M
nnFormer (arXiv 2021) [53]	45.14	42.28	36.53	53.97	45.54	60.67	24.28	32.19	149.49M
Swin-UNETR (CVPR 2022) [54]	65.54	72.04	40.57	60.05	50.26	64.32	35.21	42.94	62.19M
UNETR++ (arXiv 2022) [42]	56.49	60.04	37.25	53.59	37.13	51.99	25.36	30.68	55.70M
3D UX-Net (ICLR 2023) [43]	57.59	58.55	34.83	52.56	45.54	60.67	28.50	32.73	53.01M
SAM-B (1 pt/slice) [4]	36.30	29.86	24.01	26.74	6.71	7.63	28.83	33.63	–
Ours (1 pt/volume)	**73.78**	**83.86**	**54.09**	**76.27**	54.78	69.55	**48.35**	**63.65**	25.46M
SAM-B (3 pts/slice)) [4]	39.66	34.85	29.80	33.24	7.87	6.76	35.26	39.31	–
Ours (3 pts/volume)	**74.91**	**84.35**	**54.92**	**77.57**	56.30	70.02	**49.43**	**65.02**	25.46M
SAM-B (10 pts/slice)) [4]	40.07	34.96	30.55	32.91	8.56	5.97	39.14	42.70	–
Ours (10 pts/volume)	**75.95**	**84.92**	**57.47**	**79.62**	56.61	69.52	**49.99**	**65.67**	25.46M

6 总结与展望

本报告基于原论介绍了 3DSAM-Adapter 方法的整体思路，并展示了相关的复现过程和结果。从结果可以看出，实验基本达到了论文中的分割效果。因此，本次复现也验证了这种对视觉通用大模型进行微调以适应下游任务的方法是行之有效的探索方向。另外，也为医学图像分割带来了新的研究思路。

参考文献

[1] Gong S, Zhong Y, Ma W, et al. 3DSAM-adapter: Holistic Adaptation of SAM from 2D to 3D for Promptable Medical Image Segmentation[J]. arXiv preprint arXiv:2306.13465, 2023.

[2]Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything.arXiv:2304.02643, 2023.

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

[5] Yichi Zhang and Rushi Jiao. How segment anything model (sam) boost medical image segmentation?arXiv preprint arXiv:2305.03678, 2023.

[6] Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, et al. Segment anything model for medical images? arXiv preprint arXiv:2304.14660, 2023.

[7] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9):1–35, 2023.

[8] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In International Conference on Machine Learning, pages 2790–2799. PMLR, 2019.

[9] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition. arXiv preprint arXiv:2302.03024, 2023.

[10] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods, 18(2):203–211, 2021.