Segment Anything

摘要：

Segment Anything Model (SAM)是一种备受瞩目的通用图像分割模型，近年来在医学图像分割领域引起了广泛关注。尽管 SAM 在处理自然图像时表现卓越，但当面对医学图像，特别是那些涉及低对比度、模糊边界、复杂形状和小尺寸物体的图像时，其性能显著下降，泛化能力受到限制。本文基于 Adapting Segment Anything Model for Clinically-Friendly and Generalizable Ultrasound Image Segmentation(SAMUS)进行了一系列复现，旨在实现更低的部署成本，使其更适用于临床应用。

具体而言，在 SAM 的基础上，引入了一个并行的 CNN 分支，将局部特征注入到 ViT 编码器中，以更精准地分割医学图像。随后，结合了位置适配器和特征适配器，使 SAM 得以从自然领域迁移到医学领域，能够适应不同输入尺寸，从大尺寸输入（1024×1024）到小尺寸输入（256×256），实现更为临床友好的部署。本文在超声心动图数据集 CAMUS 上进行了一系列实验证明，实验结果显著。

关键词：SAM 医学图像分割 适配器 CNN

目录

1 引言

医学图像分割作为识别和突显医学图像中特定器官、组织和病变的关键技术，是计算机辅助诊断系统的关键组成部分，而且许多深度学习模型已经被引入，展现出巨大的应用潜力。

然而，现有的深度学习模型通常是为特定对象设计的，应用于其他对象时需要重新训练，给临床使用带来很大的不便。在这一背景下，segment anything model（SAM）作为一种通用的视觉分割基础模型，因其出色的跨不同对象的分割能力和强大的零样本泛化能力而备受好评。通过用户提供的简单提示，如点、边界框和粗掩码，SAM 能够轻松地适应各种分割应用程序。

这一模型设计的独特之处在于其能够将多个独立的医学图像分割任务整合到一个统一的框架中，形成通用模型，从而显著促进了临床应用的部署。

尽管已经构建了目前为止最大的医学图像分割数据集（即 SA1B），但由于缺乏可靠的临床注释，SAM 在医疗领域的性能却可能迅速下降。为了应对这一挑战，已有一些基础模型提出，试图通过在医学数据集上调整 SAM 来适应医学图像分割任务。然而，与 SAM 一样，它们在特征建模之前对输入图像执行无重叠的 16 倍标记化，这破坏了对识别小目标和边界至关重要的局部信息，使得这些模型难以分割具有复杂/线状形状、弱边界、小尺寸或低对比度的临床对象。此外，它们往往需要输入图像的 1024 × 1024 大小，由于生成的输入序列较长，对 GPU 的计算需求也较高，造成了较大的负担。

在本研究中，本文在 segment anything model（SAM）的基础上进行了改进，对 SAMUS 进行复现，旨在将 SAM 卓越的分割性能和强大的泛化能力引入医学图像分割领域，并在此过程中降低计算复杂度。在保持 SAM 原始框架模型不变的前提下，着重对其图像编码器进行了精心设计。

首先，通过减小输入大小来缩短 vit 分支的序列长度，以降低计算复杂度。随后，引入了特征适配器和位置适配器，以实现 vit 图像编码器在从自然域到医学域的微调过程中的优化。为了弥补 vit 图像编码器中可能缺失的局部信息，引入了一个并行的 CNN 分支图像编码器，并将其与 vit 分支同时运行。此外，引入了一个跨分支关注模块，确保 vit 分支中的每个补丁都能够充分吸收来自 CNN 分支的局部信息。

为了全面评估新模型的性能，本文主要以超声心动图 CAMUS 为基准进行了各种性能评估。旨在维持 SAM 原有优势的同时，更好地适应医学图像分割的特殊需求。

2 相关研究工作

2.1 视觉调优

随着计算机视觉领域基础模型令人瞩目的进展，人们提出了一系列视觉调优方法，以使这些基础模型更好地适应各种下游任务。一般而言，最近的可视化调优方法可以归为五大类，包括微调、参数调优、重新映射调优、提示调优和自适应调优。具体而言，微调方法涵盖了对预训练模型整个参数集的调整，以及对预训练模型特定部分的有选择性微调。参数调优方法直接修改模型参数的权重或偏差。重新映射方法通过知识蒸馏、基于权重的重新映射或基于体系结构的重新映射，将从预训练模型中学到的信息转移到下游模型。

提示调优方法通过将一组可学习参数与输入结合或设计一个子网络来生成视觉提示，引入下游任务的知识。适配器调优是目前最广泛采用的策略之一，它通过将额外的可学习参数与固定的预训练模型结合，促进下游任务的学习。这些调优方法为基础模型提供了灵活性，使其能够更好地适应不同领域和任务的需求。

2.2 基础模型

基础模型是指在广泛的数据集上进行预训练（通常采用自监督学习）的模型，能够通过微调或上下文学习机制迅速应用于下游任务。与小型模型相比，基础模型的一个区别在于，随着训练数据量和模型参数的增加，基础模型的智能化程度逐渐显现。目前，基础模型在自然语言处理（NLP）领域已取得较为成熟的进展，如 BERT、GPT-2、ToBERTa、T5 等，它们采用了自监督训练的范式，在语境理解方面表现出色。此外，像 CLIP 和 BEiT-3 等模型成功地连接了多模态之间的关系，实现了语义上的对齐，使模型更加通用化。

在计算机视觉领域，"Segmentation Anything" （SAM）首次提出了自然图像分割的基础模型，通过在 1100 万张图像上进行预训练，生成了 1 亿个掩码。该模型允许用户提供各种类型的提示，如点、框、掩码和文本，以便模型能够生成用户需要的分割掩码。

2.3 SAM 在医学图像分割的应用

尽管 SAM 在自然图像中展现出色的性能，但在某些医学图像分割任务中，尤其是在处理形状复杂、边界模糊、尺寸小或对比度低的物体时，其表现并不理想。

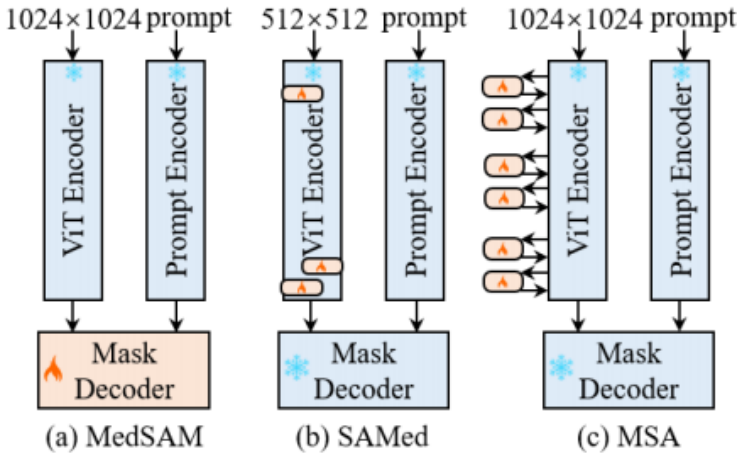


图 1 基于 SAM 的医学图像分割模型结构的比较

为了缩小 SAM 在自然图像和医学图像之间的性能差距，一些方法已经被提出，以有限的下游医学数据集对 SAM 进行调整。

MedSAM 采用了一种冻结图像编码器和提示编码器的策略，在医学图像上以可接受的成本对 SAM 进行训练，其重点是对 SAM 的掩模解码器进行调整。SAMed 引入了低秩（low-rank-based, LoRA）策略在图像编码器上进行调优，以更低的计算成本使 SAM 更适用于医学图像分割。MSA 在 vit 图像编码器的每个变压器层上采用两个上下游转换器，以引入特定于任务的信息。如图 1 所示，相较于当前基于 SAM 的基础模型，本文提出的模型更专注于补充局部特征和降低 GPU 消耗，这对于在临床场景中实现准确且易于部署的医学图像分割至关重要。

3 模型及优化

3.1 模型概述

本文复现的模型架构整体继承自 SAM(segment anything model)，如图 2 所示。SAM 有三个组件，分别是图像编码器、灵活的提示编码器和掩码解码器。接下来对这三个组件进行简单的介绍。

图像编码器：基于可扩展和强大的预训练方法，使用了 MAE 预训练的 VIT，最小限度地适用于处理高分辨率输入。在每次训练时，图像编码器对每张图像运

行一次并且在提示模型之前进行使用。

提示编码器：SAM 在进行提示时，考虑了两组提示：稀疏提示（点、框、文本）和密集提示（掩码）。通过位置编码来表示点和框，并对每个提示类型的学习嵌入和自由形式的文本与 CLIP 中现成的文本编码相加。密集的提示，即掩码，使用卷积进行嵌入并通过图像嵌入进行元素求和。

掩码解码器：掩码解码器有效地将图像嵌入、提示嵌入和输出 token 映射到掩码，设计的灵感来源于 DETR，采用了对带有动态掩膜预测头的 Transformer decoder 模块的修改。

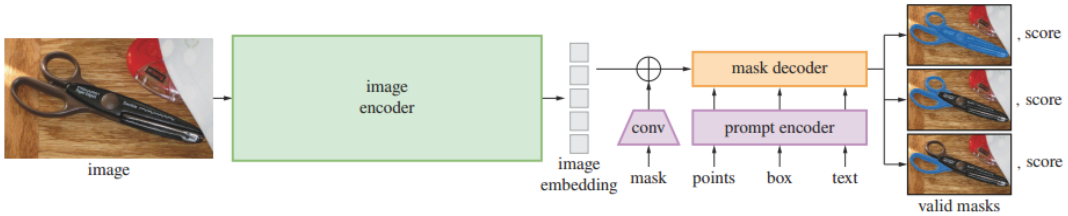


图 2 SAM 模型框架

本文复现的模型保留了提示编码器和掩码解码器的结构和参数，未做任何调整，相比之下，对图像编码器进行了精心修改，以解决局部特征不足和计算内存消耗过多的挑战，使其更适合临床友好的分割。

主要修改包括减小输入大小、向图像编码器中引入适配器、在图像编码器中增加 CNN 分支。具体来说，将原本 SAM 的输入空间分辨率 1024×1024 像素缩小到了 256×256 像素。将输入的尺寸减少大大降低了 GPU 内存成本；向图像编码器引入的适配器主要包括两种：位置适配器和功能适配器，位置适配器是为了更好的适应更短序列的全局位置嵌入，特征适配器主要是微调预训练的图像编码器；在图像编码器中引入 CNN 分支，使 CNN 分支和 VIT 分支并行，向后者提供互补的局部信息。

3.2 ViT 分支中的适配器

为了将经过训练的 SAM 图像编码器（即 ViT 分支）推广到更小的输入尺寸和医学图像域，引入了位置适配器和特征适配器。这些适配器能够有效地调整 ViT 分支，同时只需使用更少的参数。

具体而言，位置适配器负责调整位置嵌入以匹配嵌入序列的分辨率。它首先通过最大池化对位置嵌入进行下采样，步长和核大小设定为 2，以确保获得与嵌入序列相同的分辨率。然后，应用核大小为 3 × 3 的卷积运算来进一步调整位置

嵌入，有助于 ViT 分支更好地处理较小的输入。

所有特征适配器都具有相同的结构，由三个组件组成：向下线性投影、激活函数和向上线性投影。每个特征适配器的过程可以表示为：

$$\mathcal{A}(x) = \mathcal{G}(xE_d)E_u \tag{1}$$

其中$\mathcal{G}$表示 GELU 激活函数，$E_d \in R^{d\times\frac{d}{4}}$和$E_u \in R^{\frac{d}{4}\times d}$是投影矩阵，$d$是特征嵌入的维数。通过这些简单的操作，特征适配器使 ViT 分支能够更好地适应医学图像域的特征分布。这一调整策略旨在增强 SAM 在医学图像分割任务中的性能和泛化能力。

3.3 CNN 分支

CNN 分支由顺序连接的卷积池块组成。具体来说，初始输入首先经过单个卷积块的处理，然后通过三个卷积池块进行进一步处理。

接着，CNN 分支中的特征图与 ViT 分支的特征图具有相同的空间分辨率。在 CNN 分支的其余部分，单个卷积块按顺序重复了 4 次，以进一步提取特征和促进信息传递。这种结构设计旨在充分利用卷积操作的局部感知能力，有助于捕获医学图像中的细节和复杂特征。整个 CNN 分支的层次结构使其能够有效地补充 ViT 分支的全局特征提取能力，从而提高 SAM 在医学图像分割任务中的性能。

3.4 训练策略

在训练模型之前，只利用在 SA-1B 数据集上训练的 SAM 模型的权重进行初始化，以继承 SAM 的参数。其余参数则采用随机初始化。在整个训练过程中，只更新适配器和 CNN 分支的参数，而其他参数保持冻结。采用组合损失函数进行监督，该函数由 dice 损失和二元交叉熵损失组成。为了方便使用，本文仅使用最简单的正点提示符。通过在标签的前景区域随机采样一个点来模拟专家提供提示的过程。

本文使用 Adam 优化器进行训练，初始学习率设置为 0.0001，批大小为 8，训练进行了 50 个 epoch。这一训练流程旨在通过有选择性地更新特定模块的参数，从而使模型更好地适应医学图像分割任务。通过使用 SA-1B 数据集进行初始化和提示符的引导，模型的训练目标是提高其性能，以更准确地适应医学图像领域的特殊需求。

4 实验

4.1 数据集

超声心动图（Echocardiography），或称心脏超声，是评估心脏功能和结构的最广泛使用和最容易获得的成像方式。具有便携、快速实时采集、高时间分辨率等优势，并且没有电离辐射的风险，是目前心血管成像的支柱。从心力衰竭到心脏瓣膜疾病，超声心动图对于诊断许多心血管疾病来说既是必要的，也是足够的。

近年来人工智能在医学影像自动分析领域取得了重大进展。然而相比与 MRI、CT 以及病理切片，深度学习在超声医学领域方面的工作却少得多，尤其是超声心动图方面。这方面的一个主要瓶颈是缺乏公开的、有良好注释的医学图像数据。因此开放的数据库中有益于人工智能与计算机视觉在超声医学方面的广泛研究与应用。下面将介绍本文使用的超声心动图数据集 CAMUS。

CAMUS 数据集由 500 名患者的临床超声心动图检查组成，且经过匿名化处理，为了保证临床数据的真实性，既没有进行先决条件也没有进行数据选择，导致 1）一些病例难以追踪；2）不同病例图像采集设置差异很大；3）可能包括心尖五腔视图，从而产生高度异质数据集。在本文中，随机将数据集按 7：2：1 的比例划分为训练集、验证集和测试集。

4.2 实验结果及分析

为了验证模型的有效性，本文与几个基准方法进行比较，包括 U-Net、CE-Net、SwinUnet、TransUnet 和 H2former。表 1 给出了不同模型下的定量结果。

表 1 不同模型下的测试结果

Method	Dice	HD
U-Net	93.56	9.90
CE-Net	93.31	9.65
SwinUnet	91.72	12.80
TransUnet	93.54	9.60
H2former	93.44	9.79
SAM	86.74	15.84
Ours	91.45	12.99

首先，从 Dice 系数的角度看，本文的复现模型（Ours）在医学图像分割任务中表现出令人满意的性能，达到了 91.45%。这一结果相较于 SAM 的 86.74% 有了显著的提升，表明改进后的模型在处理医学图像时能够更准确地捕捉目标结构。

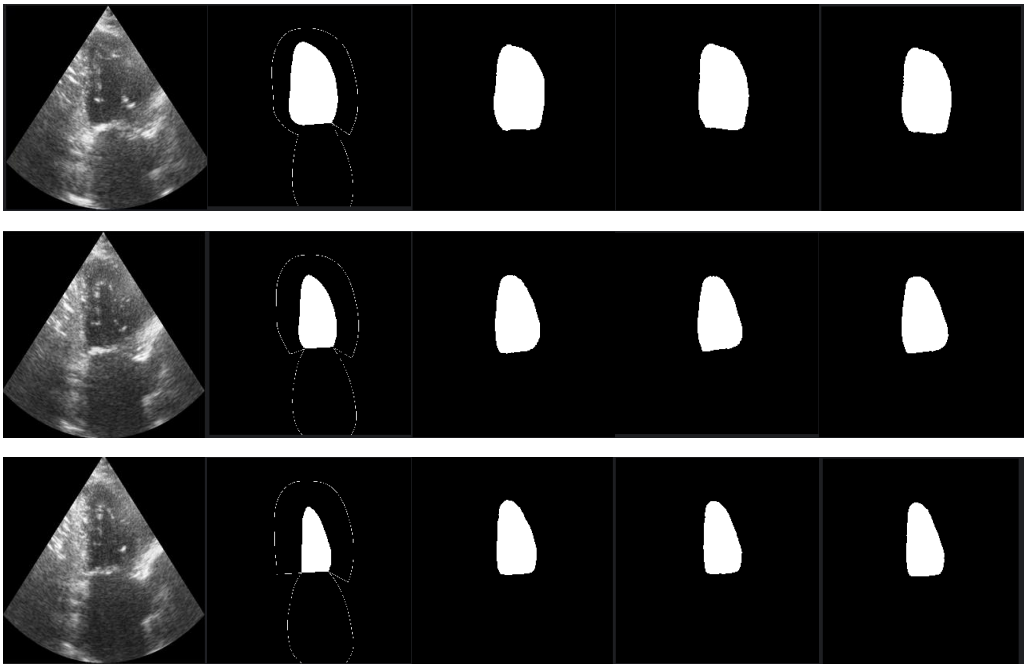
其次，对比 HD 指标，我们的复现模型（Ours）在 12.99 的 HD 值上表现较好，相较于 SAM 的 15.84 有明显改善。这说明我们的方法在边界定义上更为准确，降低了分割结果的空间误差。这说明我们的复现达到我们的预期，改进是有效的。

表 2 基于 SAM 的其他模型同本文模型的对比

Method	Dice	HD
MedSAM	92.61	11.72
SAMed	92.87	11.34
MSA	90.8	12.6
Ours	91.45	12.99

其次为了表明复现后的模型相比于其他基于 SAM 模型有优越的性能，本文同样和同类的方法进行了对比，结果如表 2 所示。可以看出本文的方法相比较之下仍然具有卓越的性能，相较于 MSA 提高了 0.65%，虽然相较于 MedSAM 和 SAMed 仍然具有一定的差距。

如图 3 所示，给出了几种模型下的可视化结果。



Input Ground Truth Ours U-Net MSA

图 3 不同模型下可视化结果

4.3 消融实验

为了验证本文模型两个模块的对模型性能的有效性，本文做了相应的消融实验。实验过程很简单，分别测试每个模块下模型性能的好坏，消融实验结果如表

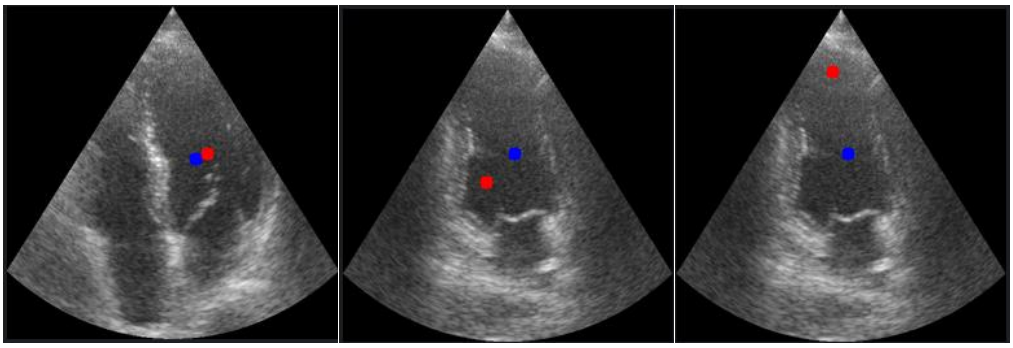
3 所示。

表 3 不同组件对模型性能的影响

Components		CAMUS	
CNN Branch	Adapter	Dice	HD
✓	×	91.03	13.2
×	✓	90.6	13.36
✓	✓	91.45	12.99

通过消融实验可以看出，两个改进均对模型的提升有帮助。CNN 分支比适配器有更显著的性能提升，但即使是一个简单的适配器对模型也有较为明显的提升。

其次，因为 SAM 是一个提示分割模型，为了探究点提示的数量和位置对模型的影响，本文同样做了一些实验，同样是在 CAMUS 数据集上，结果如表 4 所示。一般来说，SAM 对点的位置和数量具有鲁棒性，对于类内变化较大的对象，不同位置的点提示对模型的性能对于 Dice 系数影响较大；此外，引入多个点提示可以提高模型的性能。



Random Entropy Max_dist

图 4 不同策略下点提示的选择

这里选取了三种点提示选择的策略，在第一个点都是随机选取的基础之上，第二个点的选取分别是随机选取、最大熵选取和最大距离选取，选取点的结果如图 4 所示。其中，蓝色点表示第一次选择的点，红色点表示第二次选择的点。本文只做了两个点提示和单点提示下的对比结果以及不同点提示选取策略下两个点提示模型性能的对比，结果如表 4 所示。通过表中的第一行和第二行，可以看出两个点提示下比单点提示下，Dice 系数上升了 0.15%；第二行到第四行的结果对比是不同点提示策略下模型性能的对比，可以看出在都是两个点提示的情况下，选取距离初始点距离最大的点有更高的提升，达到了 0.31%。

表 4 不同点提示策略下对模型性能的影响

Method	Dice	HD
Ours	91.45	12.99
Ours+Random	91.6	12.87
Ours+Entropy	91.53	12.91
Ours+Max_dist	91.76	12.71

5 总结

在本文中，基于 SAM 模型框架做出了一些复现，用于临床友好的医学图像分割。具体来说，在 SAM 模型框架的基础之上，在图像编码器增加了一个并行的 CNN 分支，并在编码器中添加了适配器，使其更好的适应于医学图像分割。以丰富的小尺寸和边界区域的特征，同时降低 GPU 的消耗。通过实验结果可以看出，本文的复现模型虽然距离一些 SOTA 模型还有一定的差距，但是改进效果是显而易见，模型的性能仍然具有较为出色的分割能力，仍然具有很大的改进空间。

其次，在消融实验中可以看出，提示点的位置和提示点的位置对模型性能也有一定的影响，因此选择一个更加优良的点提示策略至关重要。

在接下来的工作中，将继续本文的实验，尝试提出新的方法，以提高模型的性能，使其达到 SOTA 水平的性能。主要聚焦于两点：1）扩展到视频分割，在其基础之上加入时空注意模块。2）尝试新的模块及方法，主要对图像编码器进行调整，以提高模型的性能。

参考文献

[1] Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; and Wang, M. 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In European Conference on Computer Vision, 205–218. Springer.

[2] Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A. L.; and Zhou, Y. 2021. Transunet: Transformers make strong encoders for medical image segmentation. arXiv:2102.04306.

[3] Gu, Z.; Cheng, J.; Fu, H.; Zhou, K.; Hao, H.; Zhao, Y.; Zhang, T.; Gao, S.; and Liu, J. 2019. Ce-net: Context encoder network for 2d medical image segmentation. IEEE Transactions on Medical Imaging, 38(10): 2281–2292.

[4] He, A.; Wang, K.; Li, T.; Du, C.; Xia, S.; and Fu, H. 2023. H2Former: An Efficient Hierarchical Hybrid Transformer for Medical Image Segmentation. IEEE Transactions on Medical Imaging.

[5] Huang, X.; Deng, Z.; Li, D.; Yuan, X.; and Fu, Y. 2022. Missformer: An effective transformer for 2d medical image segmentation. IEEE Transactions on Medical Imaging.

[6] Huang, Y.; Yang, X.; Liu, L.; Zhou, H.; Chang, A.; Zhou, X.; Chen, R.; Yu, J.; Chen, J.; Chen, C.; et al. 2023. Segment anything model for medical images? arXiv:2304.14660.

[7] Kiranyaz, S.; Degerli, A.; Hamid, T.; Mazhar, R.; Ahmed, R. E. F.; Abouhasera, R.; Zabihi, M.; Malik, J.; Hamila, R.; and Gabbouj, M. 2020. Left ventricular wall motion estimation by active polynomials for acute myocardial infarction detection. IEEE Access, 8: 210301–210317.

[8] Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.- Y.; et al. 2023. Segment anything. arXiv:2304.02643.

[9] Leclerc, S.; Smistad, E.; Pedrosa, J.; Østvik, A.; Cervenansky, F.; Espinosa, F.; Espeland, T.; Berg, E. A. R.; Jodoin, P.- M.; Grenier, T.; et al. 2019. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. IEEE transactions on medical imaging, 38(9): 2198–2210.

[10] Lian, D.; Zhou, D.; Feng, J.; and Wang, X. 2022. Scaling & shifting your features: A new baseline for efficient model tuning. In Advances in Neural Information Processing Systems, 109–123.

[11] Liu, X.; Song, L.; Liu, S.; and Zhang, Y. 2021. A review of deep-learning-based medical image segmentation methods. Sustainability, 13(3): 1224.

[12] Ma, J.; and Wang, B. 2023. Segment anything in medical images. arXiv:2304.12306.

[13] Pedraza, L.; Vargas, C.; Narvaez, F.; Dur ´ an, O.; Mu ´ noz, ˜ E.; and Romero, E. 2015. An open access thyroid ultrasound image database. In 10th International Symposium on Medical Information Processing and Analysis, volume 9287, 188–193. SPIE.

[14] Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image

Computing and Computer-Assisted Intervention, 234–241. Springer.
[15] Wang, H.; Chang, J.; Luo, X.; Sun, J.; Lin, Z.; and Tian, Q. 2023. Lion: Implicit vision prompt tuning. arXiv:2303.09992.
[16] Wu, H.; Chen, S.; Chen, G.; Wang, W.; Lei, B.; and Wen, Z. 2022. FAT-Net: Feature adaptive transformers for automated skin lesion segmentation. Medical Image Analysis, 76: 102327.
[17] Wu, J.; Fu, R.; Fang, H.; Liu, Y.; Wang, Z.; Xu, Y.; Jin, Y.; and Arbel, T. 2023. Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv:2304.12620.
[18] Wunderling, T.; Golla, B.; Poudel, P.; Arens, C.; Friebe, M.; and Hansen, C. 2017. Comparison of thyroid segmentation techniques for 3D ultrasound. In Medical Imaging, volume 10133, 346–352. SPIE.