基于 SAM 目标分割算法在多模态图像融合上的应用

摘要 此次复现主要介绍一个可分割任何图像目标的算法 Segment Anything model(SAM)。SAM 由 meta 公司提出，作为第一款图像分割通用大模型，该模型的训练通过建立了迄今为止最大的分割数据集，在 1100 万张图像和超过 10 亿个掩模所训练而成。该模型能够利用图像的语义信息来对图像进行分割，并标记出不同区域的内容。这一技术在计算机视觉和自然语言处理等领域具有广泛应用前景。通过分析已有的 segment anything model 研究成果，利用其强大的分割效果，在不同模态下对特定目标进行分割，所得的分割结果作为多模态配准的基准，实验证明，该方法能高效完成 CT 和 X 光影像的配准，有较好的工程应用价值。

关键字：SAM；分割；配准；融合；

1 引言

在网络规模的数据集上预训练的大型语言模型具有强大的零采样和少采样泛化能力，正在彻底改变。这些基础模型可以推广到训练期间看到的任务和数据分布之外的任务和数据分布。此功能通常通过提示工程实现，其中使用手工编写的文本来提示语言模型为手头的任务生成有效的文本响应。当使用来自网络的大量文本语料库进行缩放和训练时，这些模型的零和少镜头性能与在某些情况下匹配微调模型。经验趋势表明，这种行为随着模型规模、数据集大小和总训练计算而

改善。

基础模型也在计算机视觉中进行了探索，尽管程度较低。也许最突出的插图是对齐来自网络的配对文本和图像。例如，CLIP 和 ALIGN b 使用对比学习来训练对齐两种模式的文本和图像编码器。经过训练后，工程文本提示可以对新的视觉概念和数据分布进行零概率泛化。这种编码器还可以与其他模块有效组合，以实现下游任务，例如图像生成。虽然在视觉和语言编码器方面已经取得了很大的进展，但计算机视觉包括超出这个范围的广泛问题，并且对于其中的许多问题，没有丰富的训练数据。

在这项工作中，我们的目标是建立一个图像分割的基础模型。也就是说，我们寻求开发一个提示模型，并使用一个能够实现强大泛化的任务在广泛的数据集上对其进行预训练。有了这个模型，我们的目标是使用即时工程解决新数据分布上的一系列下游分割问题。

2 相关工作

2.1 交互式分割

图像语义分割：研究利用用户提供的粗糙标签或绘制的边界来辅助图像语义分割的方法。这些方法通常将用户的标签与图像特征相结合，从而实现更准确的分割结果。交互式视频分割：研究利用用户提供的标签或标记的视频帧来实时分割视频中的目标对象。这种方法可以应用于视频编辑、视觉特效等领域。用户交互：研究如何最大限度地利用用户提供的交互信息来改善分割结果。例如，利用用户提供的

分割标记来指导和优化分割算法，从而减少用户的交互成本。交互式3D分割：研究如何利用用户提供的3D标记或交互来实现更准确的三维模型分割。这种方法可以应用于计算机辅助设计、医学图像分析等领域。强化学习与交互式分割：研究利用强化学习算法来指导分割过程，根据用户的反馈不断优化分割算法。这种方法可以实现自动化的分割过程，并根据用户的需求进行自适应调整。

2.2 轮廓检测

基于深度学习的边缘检测：利用深度神经网络，如卷积神经网络（CNN），设计和训练用于边缘检测的模型。这些模型能够学习到图像中丰富的边缘特征，并实现更准确的边缘检测结果。多尺度边缘检测：研究如何在不同的尺度下检测图像中的边缘。这种方法可以捕捉到不同粗细的边缘，并提高对区域边缘的检测准确性。基于超像素的边缘检测：利用超像素分割技术对图像进行预处理，然后在超像素上进行边缘检测。这种方法可以减少计算复杂度，并提高边缘检测的鲁棒性。基于概率图模型的边缘检测：利用概率图模型，如马尔可夫随机场（MRF）或条件随机场（CRF），对局部像素进行建模，并进行全局优化。这种方法可以将上下文信息纳入边缘检测过程，提高检测结果的准确性。基于边缘增强的边缘检测：研究如何通过增强边缘信号来提高边缘检测的结果。例如，使用边缘增强算法（如Canny边缘检测）或结合其他图像处理技术（如锐化滤波）进行边缘检测。

2.3 语义分割

基于深度学习的语义分割：深度学习方法，如卷积神经网络（CNN）和全卷积网络（FCN），在语义分割领域取得了巨大进展。这些方法能够端到端地学习像素级别的分类和定位，从而实现高精度的语义分割。多尺度语义分割：研究者们致力于设计能够在不同尺度下对图像进行语义分割的方法，以捕捉不同尺度下的语义信息，并提高分割的准确性。语义分割与实例分割的结合：研究者们开始探索如何将语义分割与实例分割相结合，以区分同一类别中不同对象的实例。这种方法可以进一步提高语义分割的效果，使得分割结果更具有语义和实例的区分度。跨模态语义分割：研究如何将不同模态的数据（如光学图像和激光雷达数据）进行融合，从而实现多模态的语义分割。这种方法对于自动驾驶等领域具有重要的应用意义。弱监督和自监督学习：研究如何利用弱监督信息（如像素级标签、图像级标签）或自监督学习来改进语义分割的性能，降低对大量标注数据的依赖。

3 本文方法

3.1 本文方法概述

使用 MAE 自监督训练得到的 ViT 模型作为图像特征的抽取网络，用来对输入图像做 embedding，参数量比较大，但是对同一张输入图像只需要计算一次 embedding，SAM 算法支持的几种 prompt 方式，对于 points、box 位置编码的形式嵌入到网络中，对于 mask 则采用几层卷积的方式提取 embedding，对于 text 则是采用 CLIP 中的文本编码

器得到对应的 embedding。本文只使用 box 和 points，对于不同的 prompt 可以重复使用 image embedding，从而降低推理压力。根据前两步得到的 image embedding 和 prompt embedding 生成有效的多个 mask 和 每个 mask 对应的置信分数。Prompt encoder 和 Mask decoder 都是轻量级的结构，参数量较少，可以在 web 端快速推理。探索无外标志物作为参考系的非刚性配准算法，本研究拟采用基于金字塔网络对多模态医学数据进行融合。金字塔网络除了强大的特征提取能力外还克服了传统 CNN 输入图像大小必须固定的问题，从而可以使得输入图像宽高比和大小任意，因此非常适合用于多模态的特征融合。总体拟采用如图 3-1 所示的融合方式: 利用金字塔变换构造了输入图像的顺序，并进行下采样对图像进行分解，得到图像的局部特征后通过计算对应的局部等效能量来评估局部区域的匹配度，最终对融合金字塔进行反变化，便可得到融合图像。

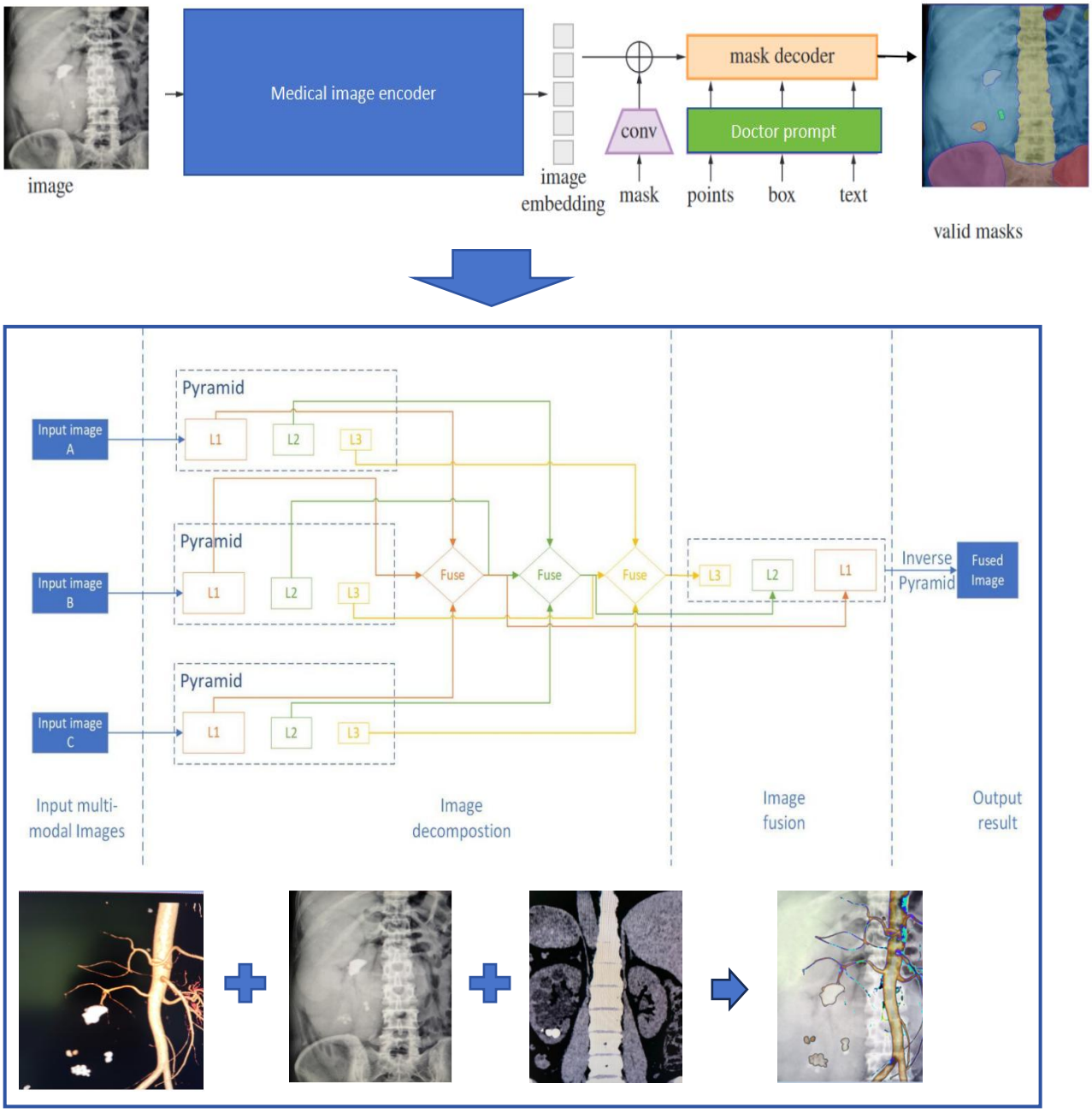


图 3-1 多模态三层金字塔融合图像网络设计

3.2 特征提取模块

输入分辨率和图像嵌入：本文采用受标准实践的启发，使用 1024×1024 的输入分辨率，并通过对图像进行缩放和在较短的一侧进行填充来获得。因此，图像嵌入为 64×64。为了减少通道维数，根据[62]的做法，作者们使用 1×1 卷积将通道减少到 256 个，然后跟随着一个同样有 256 个通道的 3×3 卷积。每个卷积后都接一个层归一化操作。提示编码器：稀疏提示通过映射到 256 维向量嵌入进行表示。点提示被表示为点位置的位置编码之和与两个学习嵌入之一的和，这两个学习嵌入表示点分别属于前景或背景。盒子提示被表示为一个嵌入对：(1)它的左上角位置编码与表示“左上角”的学习嵌入之和，以及(2)使用表示“右下角”的学习嵌入时相同的结构。综上所述，本文的特征提取方法通过特定的输入分辨率和图像嵌入的设置，以及灵活的提示编码器，从而实现对视觉和非视觉信息的高效提取和表示。

3.3 损失函数定义

本次复现主要使用三种 loss 函数，focal loss 、 dice loss 和 Match_Loss ，Focal loss 是最初由何恺明提出的，最初用于图像领域解决数据不平衡造成的模型性能问题。本文试图从交叉熵损失函数出发，分析数据不平衡问题，focal loss 与交叉熵损失函数的对比，给出 focal loss 有效性的解释。Focal loss 的定义：

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t)$$，(1-pt)ˆγ为调变因子，这里γ≥0，称为聚焦参数；Dice Loss，也叫Soft Dice Coefficient，是一种用于图像分割任务的损失函数。它基于目标分割图像与模型输出结果之间的重叠区域的比例计算出分数。与交叉熵损失函数相比，它更适合于处理难分割的目标。Dice Loss的公式如下：

$$s = \frac{2|X \cap Y|}{|X| + |Y|} = \frac{2TP}{2TP + FN + FP}$$

；最后配准的约束是利用不同模态的图像的同一个目标的轮廓作为约束条件，起定义为：Match_Loss = |Sobel(CT) - Sobel(X)|/N。

4 复现细节

4.1 与已有开源代码比：

1、把SAM模型移植到标注工具上，实现半自动快速数据标注。

2、把SAM的分割结果应用到个人数据集上并实现多模态融合配准算法Match_Loss = |Sobel(CT) - Sobel(X)|/N

4.2 实验环境搭建

1、系统平台：win10

2、硬件环境：intel i7 , nvidia V100 , ram 32G

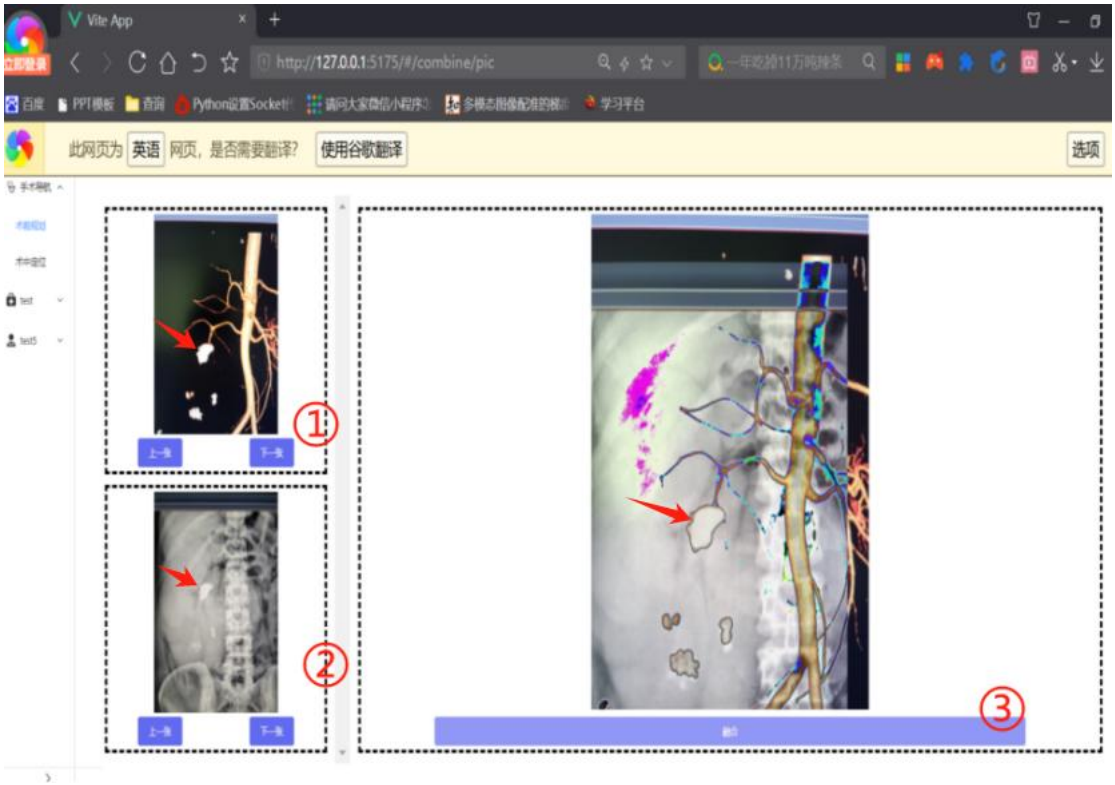
3、extras_require={

"all": ["matplotlib", "pycocotools", "opencv-python",

"onnx", "onnxruntime","python3.10","pytorch1.10.1","cuda 11.3",
"VUE"],
"dev": ["flake8", "isort", "black", "mypy"],
},

4.3 界面分析与使用说明

如图 4-1，启动前端后，①和②可选择 CT 和 X 光两模态的图，点击③处的融合后，SAM 算法被调用进行不同模态下的特定目标的分割，分割结果作为配准的基准，进行 CT 和 X 光的融合，如下图箭头指向的是病灶肾结石作为特定的分割和配准目标，③处输出为融合结果。



4.4 创新点

使用 SAM 实现多模态医学影像的深度学习融合算法在肾结石的分割和融合的应用研究，搭建 CT、X 光、超声等医学影像的目标识别模型和分割模型，并设计一种先验知识和 prompts 模式的非刚性多模态配准算法，无需外置标志物，为手术导航系统提供一种新型的关键技术。

5、实验结果分析

分别对 131 张肾结石 CT 图片和 131 张 X 光肾结石图片进行测试，SAM 在 X 光数据分割效果欠佳，实验结果如下表 5-1 和 5-2。

Table 5-1 对比 SAM 与 yolov5 在 ct 上的分割效果

COMPARATIVE ANALYSIS OF TWO DIFFERENT ALGORITHMS ON CT				
	Precision	Recall	F1 score	Accuracy(%)
SAM	0.962	0.954	0.958	0.95
yolov5	0.933	0.896	0.914	0.899

Table 5-2 对比 SAM 与 yolov5 在 x-ray 上的分割效果

COMPARATIVE ANALYSIS OF TWO DIFFERENT ALGORITHMS ON X-Ray				
	Precision	Recall	F1 score	Accuracy(%)
SAM	0.853	0.804	0.863	0.813
yolov5	0.921	0.811	0.905	0.875

6 总结与展望

在过去的研究中，SAM 已经取得了一定的进展。这种模型可以通过学习图像的语义信息，将图像分割成不同的区域，并对每个区域进行标记。这种模型可以被应用于许多领域，如计算机视觉、自然语言

处理等。研究者们已经提出了许多不同的SAM模型，并且在各种数据集上取得了显著的效果。

未来的研究可以进一步改进SAM的准确性和稳定性，使其能够在更复杂的场景下表现出色。另外，研究者们可以尝试结合不同的模型和技术，以进一步提高SAM的性能。在医学领域，不妨研究基于超声视频的SAM动态跟着分割。最终，希望通过不断的研究和开发，使segment anything model成为一个更加成熟和实用的技术。

参考文献

[1] Kirillov A, Mintun E, Ravi N, et al. Segment anything[J]. arXiv preprint arXiv:2304.02643, 2023.

[2] Huang Y, Yang X, Liu L, et al. Segment anything model for medical images?[J]. Medical Image Analysis, 2023: 103061.