EMCAD：用于医学图像分割的高效多尺度卷积注意力解码器

摘要

在医学图像分割中，特别是计算资源有限的场景下，一个高效且有效的解码机制至关重要。但是，目前的大部分解码机制都伴随着昂贵的计算成本。为了解决这一问题，本文工作聚焦于复现一个由 Rahman 等人提出的 EMCAD，一种新的高效多尺度卷积注意力解码器，旨在优化性能和计算效率。EMCAD 利用独特的多尺度深度卷积块，通过多尺度卷积显著增强特征图。EMCAD 还采用了通道、空间和分组门控注意力机制，这些机制在捕获复杂空间关系和聚焦显著区域方面非常有效。通过使用组卷积和深度卷积，EMCAD 非常高效且具有良好的扩展性。本工作在六个医学图像分割任务的 12 个数据集上进行了严格的评估，结果表明 EMCAD 实现了先进的性能，同时在 Params 和 FLOPs 上分别减少了 79.4% 和 80.3%。此外，EMCAD 对不同编码器的适应性和在分割任务中的多功能性，进一步确立了其作为一种有前景的工具。

关键词：医学图像分割；多尺度卷积注意力解码器；深度学习

1 引言

在医学诊断和治疗策略中，自动化医学图像分割对于识别和分类图像中的关键区域（如病变、肿瘤或整个器官）至关重要，如图1是对腹部的多器官进行分割的方法对比。在医学图像分割领域，特别是在计算资源受限的场景下，对于高效且有效的解码机制的需求更为迫切。尽管已有多种基于 U 形卷积神经网络（CNN）架构和注意力机制的模型在这一领域取得了高质量的分割结果，但这些模型通常计算成本昂贵，这限制了他们在实际应用中的可行性。最近，视觉变换器（ViTs）在医学图像分割任务中显示出了捕捉像素间长距离依赖的潜力，但在理解局部空间上下文方面存在不足，需要借助局部卷积注意力来更好的把握空间细节，这也会导致对计算资源的要求较高。

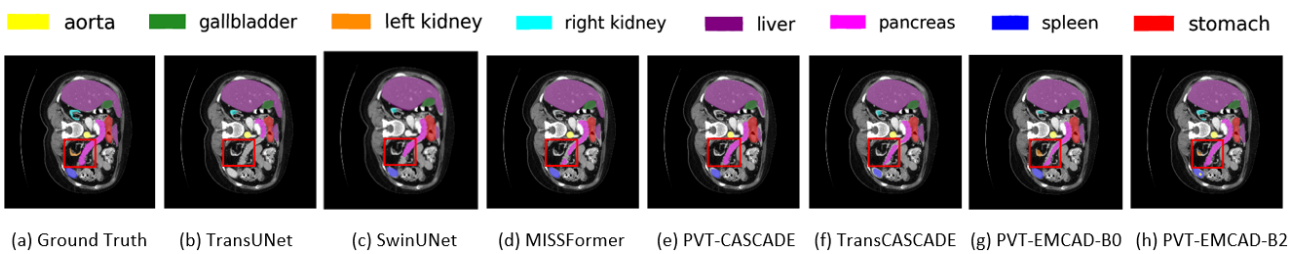

图 1. 多器官分割图实例

鉴于以上挑战，本文引入了 EMCAD，一种新的高效多尺度卷积注意力解码器。EMCAD 通过高效的多尺度卷积增强了特征图，同时通过使用通道、空间和分组（大核）门控注意力机制来整合复杂的空间关系和局部注意力。其中，EMCAD 由高效的多尺度卷积注意模块（MSCAM）组成，用于稳健地增强特征图，大内核分组注意门（LGAG）用于通过门控注意机制与跳跃连接融合来细化特征图，用于上采样的高效上卷积块（EUCB），以及产生分割输出的分割头（SH）。总的来说，本文的 EMCD 由以下几个部分组成：

（1）多尺度卷积注意力模块（MSCAM）：通过深度卷积在不同尺度上对特征图进行处理，增强特征表示并一直无关区域。

（2）大核分组注意力门（LGAG）：通过分组卷积结合大核，进一步融合编码器和解码器之间的特征，以捕捉更大的上下文信息。

（3）高效上采样卷积块（EUCB）：用于逐步上采样特征图，将特征的空间分辨率提升到目标输出的分辨率。

（4）分割头（SH）：每个阶段的特征图都通过 1x1 卷积输出一个分割图，最终整合多个分割图生成最终的分割结果。

本文后续章节按照以下逻辑展开，第二章主要介绍与本文相关的工作以及这些工作的优缺点；第三章主要详细介绍本文方法框架；第四章主要介绍复现的相关细节，包括但不限于源码来源、环境搭建、数据准备等；第五章主要分析复现结果以及相关的实验；第六章总结本文以及后续工作的期望。

2 相关工作

2.1 视觉编码器

传统卷积神经网络（CNNs），如 AlexNet [13]、VGG [21] 和 ResNet [10] 等，由于其熟练的使用图像中的空间关系，已经成为编码器的基础。其中，AlexNet 和 VGG 利用深度卷积层逐步提取特征；ResNet 引入了残差间距，通过解决梯度消失来训练层数更多的网络 GoogleNet [22] 引入了初始模式，允许更有效的计算各种尺度的表示；MobileNets [11] 通过轻量级的深度可分离卷积将 CNN 用于移动设备中；EfficientNet [23] 通过复合缩放为 CNN 引入了可扩展的架构设计。虽然 CNN 在许多视觉应用中都举足轻重，但是由于受到局部限制，通常缺乏捕捉图像内长距离依赖关系的能力。

由 Dosovitskiy [8] 等人首次提出的视觉转换器（ViTs）利用自我注意力（SA）实现了像素间的远距离关系的学习。此后，通过整合 CNN、SA 和引入新的框架，ViTs 不断得到增强。Swin Transformer [15] 融合了滑动窗口注意机制，而 SegFormer [28] 利用 MixFFN 实现了分层结构。PVT [25] 使用空间缩减注意力，并在 PVTv2 [26] 中通过重叠补丁嵌入和线性复杂度注意层加以完善。MaxVit [24] 引入了多轴自注意力，形成了分层 CNN 变换解码器。虽然 Vit 解决了 CNN 在捕捉长距离像素依赖关系方面的局限性，但是他们在捕捉橡塑件的局部空间关系方面面临着较大的挑战。

2.2 医学图像分割

医学图像分割涉及像素分类，以识别不同成像模式（如内窥镜、核磁共振成像或 CT 扫描）中的各种解剖结构，如病变、肿瘤或器官。UNet [20] 因其简单而有效的编码—解码器设计而特别受青睐。UNet 率先采用这种方法，利用跳跃连接来融合不同分辨率阶段的特征。UNet++ [29] 则是将嵌套的编码—解码器路径与密集的跳跃连接结合在一起。UNet3+ [12] 在这些理念的基础上进行了扩展，引入了全面的跳跃路径，促进了全面的特征融合。DC-UNet [17] 则取得了进一步的优化，它在跳跃连接中集成了多分辨率卷积方案和残差路径。DeepLab 系列，包括 DeepLabv3 [5] 和 DeepLabv3+ [6]，引入了无差别卷积和空间金字塔池处理多尺度信息。SegNet [1] 使用池化指数对特征图进行上采样，保留边界细节。总之，以上的 U 型模型已成为医学图像分割领域的基准。

视觉变换器利用其捕捉全局像素关系的能力，成为医学图像分割领域的一股强大力量。TransUNet [3] 将用于局部特征提取的 CNN 和用于全局上下文的变换器进行了新颖的融合，增强了局部和全局特征捕捉能力。Swin-Unet [2] 在此基础上进行了扩展，将 Swin 变换器块纳入 U 型模型，用于编码和解码过程。在以上工作的基础上，MERIT [18] 引入了多尺度分层变换器，在不同的窗口大小中采用 SA，从而增强了模型捕捉对医学图像分割至关重要的多尺度特征的能力。

有人研究了将注意力机制整合到 CNN 和基于 Transformer 的系统中，以增强医学图像分割能力。PraNet [9] 采用反向注意力策略进行特征细化。PolypPVT [7] 利用 PVTv2 [26] 作为主干编码器，并在其解码阶段加入了 CBAM [27]。CASCADE [19] 是一种新颖的级联解码器，它结合了信道和空间注意力，在多个阶段完善从 Transformer 编码器中提取的特征，最终实现高分辨率的分割输出。虽然 CASCADE 通过整合来自 Transformers 的局部和全局洞察力，在分割医学图像方面取得了显著的性能，但由于在每个解码阶段都使用了三重 3×3 的卷积层，因此其计算效率较低。

为解决以上的局限性，本文提出 EMCAD，一种新的高效多尺度联级卷积注意力解码器。该解码器可以利用多尺度卷积注意力模块完善特征图并纳入局部注意力，同时由于此解码器采用多尺度卷积块、大内核分组注意力门和高效的上卷积块，因此减少了参数和计算量，降低了对计算资源的要求。

3 本文方法

3.1 EMCAD 整体框架概述

此部分是对本文将要复现的工作进行概述，其模型如图 2所示：

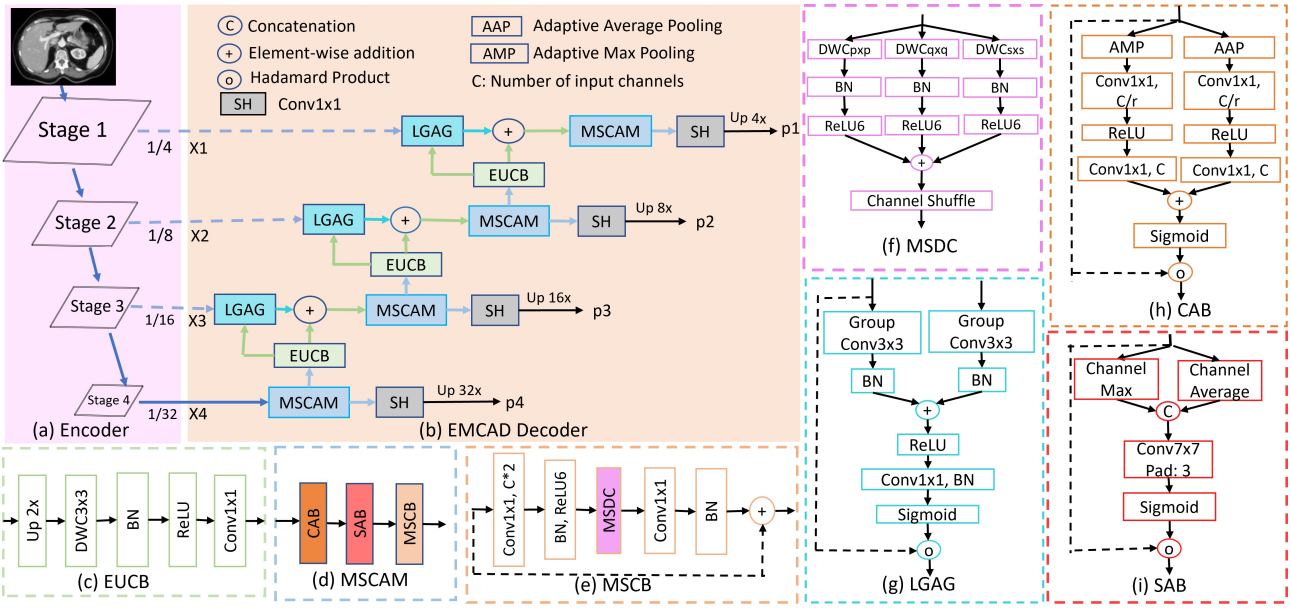

图 2. 高效的多尺度卷积注意力解码器整体架构

该模型由一个分层编码器和高效多尺度卷积注意力解码器（EMCAD）组成，通过编码器提取不同尺度的特征图，然后利用 EMCAD 中的多尺度卷积注意力模块（MSCAM）、大核分组注意力门（LGAG）和上采样模块（EUCB）逐步融合和增强这些特征，实现精确的医学图像分割。具体而言，本文使用了四个 MSCAM 来完善从编码器四个阶段中提取的金字塔特征（即图 2中的 X1、X2、X3、X4）。在每个 MSCAM 之后，使用 SH 生成该阶段的分割图，然后使用 EUCB 将细化后的特征图放大，并添加到相应的 LGAG 的输出中。最后合成四个不同的分割图，生成输出高分辨率的分割结果，模型在保持高精度的同时显著降低了计算成本。接下来详细介绍解码器的不同模块。

3.1.1 多尺度卷积注意力模块（MSCAM）

本文通过引入一个有效地多尺度卷积注意模块来细化特征映射。MSCAM 由一个通道注意块 (CAB(.))、一个空间注意块 (SAB(.)) 和一个高效的多尺度卷积模块 (MSCB(.)) 来组成增强保留上下文关系的特征映射。其中，CAB(.) 用于强调相关通道，SAB(.) 用于捕捉局部上下文信息，MSCB(.) 用于增强保留上文关系的特征映射。MSCAM(.) 如公式 1 所示：

$$MSCAM(x) = MSCB(SAB(CAB(x))) \tag{1}$$

其中 x 是输入张量。由于在多个尺度上使用深度卷积，本文的 MSCAM 比（标记）中提出的卷积注意模块（CAM）更有效，计算成本明显更低。

MSCB(.) 是本文引入的一种高效的多尺度卷积块用来增强级联扩展路径生成的特征，此模块遵循 MobileNetV2 的反向残差（IRB）的设计。与 IRB 不同的是，本文的 MSCB 在多个尺度上执行深度卷积，并使用通道洗牌来跨组洗牌通道。MSCA(.) 如公式 2 所示：

$$MSCB(x) = BN(PWC_2(CS(MSDC(R_6(BN)(PWC_1(x)))))) \tag{2}$$

MSCAM 模块中使用通道注意块（CAB(.)）为每个通道分配不同程度的重要性，从而强调更多的相关特征，同时抑制不太有用的特征。基本上，CAB(.) 可以确定需要关注哪些特征

映射并细化它们。CAB(.) 如公式 3 所示：

$$CAB(x) = \sigma(C_2(R(C_1(P_m(x)))) + C_2(R(C-1()P_\alpha(x)))) \circledast x \tag{3}$$

MSCAM 模块中使用的空间注意块 SAB(.) 通过空间注意力来模仿人脑的注意力过程，将注意力集中在输出图像的特定部分。SAB(.) 会确定在特征图中的重点位置，然后增强这些特征，这一过程增强了模型识别和相应相关空间特征的能力，对图像分割极为重要。SAB(.) 如式 4 所示：

$$SAB(x) = \sigma(LKC([Ch_max(x), Ch_avg(x)])) \circledast x \tag{4}$$

3.1.2 大核分组注意力门（LGAG）

MSCAM 模块中引入了一种新的大内核分组注意力门（LGAG），用以逐步将特征图与注意力系数相结合，网络通过学些这些系数来允许更高程度地激活相关特征并抑制不相关特征。此过程采用从高级特征派生的门控信号来控制网络不同阶段的信息流，从而提高医学图像分割的精度。在本文的 $q_{att}(.)$ 函数中，通过应用单独的 3×3 组卷积 $GC_g(.)$ 和 $GC_x(.)$ 来处理门控信号 g 和输入特征图 x。生成的特征图通过 ReLU(R(.)) 层激活，然后应用 1×1 卷积 (C(.)) 和 BN(.) 层来获得单通道特征图。最后将生成的单通道特征图传递给 Sigmoid($\sigma(.)$) 激活函数以产生注意力系数。该转换的输出用于通过元素乘法缩放输入特征 x，产生注意力门控特征 LGAG(g, x)。LGAG(.) 用公式 5、6 表示：

$$q_{att}(g, x) = R(BN(GC_g(g) + BN(GC_X(x)))) \tag{5}$$

$$LGAG(g, x) = x \circledast \alpha(BN(C(q_{att}(g, x)))) \tag{6}$$

3.1.3 高效的上卷积块（EUCB）

EMCAD 中使用了一个高效上卷积块对当前阶段的特征图进行逐步的上采样，以匹配下一个跳跃连接的特征图的维度和分辨率。EUCB 先使用了比例因子为 2 的 $U_p(.)$ 来放大特征图，然后通过使用深度卷积 DWC(.)，在通过使用 BN(.) 和 ReLU(.) 激活来增强放大的特征图。最后通过使用 1×1 卷积来减少通道以与下一阶段匹配。由于使用的是深度卷积而不是 3×3 卷积，因此 EUCB 效率会很高。EUCB(.) 如式 7 所示：

$$EUCB(x) = C_{1\times1}(ReLU(BN(DWC(U_p(x))))) \tag{7}$$

3.1.4 分割头（SH）

此文使用分割头从解码器四个阶段的细化特征图产生分割输出。SH 层将 1×1 卷积 $Conv_{1\times1}(.)$ 应用于具有 ch_i 通道的细化特征图，并产生输出。

$$SH(x) = Conv_{1\times1}(x) \tag{8}$$

3.2 多级损失聚集和输出聚合

3.2.1 多级损失聚集

EMCAD 解码器的四个分割图在四个阶段分别产生四个预测图 P1、P2、P3 和 P4，其中 P4 是解码器的最终输出图像。受 MERIT 多分类分割工作的启发，本文采用了一种名为 MUTATION 的损失组合方法。这包括计算来自 4 个头部得到所有可能预测组合的损失共计 15 个独特的预测，然后讲这些损失相加。本文的重点是在训练过程中尽量减少这种累计组合的损失。对于二元分割，使用附加项 $L_{p1+p2+p3+p4}$ 优化加性损失，如公式 9 所示：

$$L_{total} = \alpha L_{p1} + \beta L_{p2} + \gamma L_{p3} + \zeta L_{p4} + \delta L_{p1+p2+p3+p4} \tag{9}$$

其中，L_{p1}、L_{p2}、L_{p3} 和 L_{p4} 为每个单独的预测图的损失，且 $\alpha = \beta = \gamma = \zeta = \delta = 1$ 是分配给每个损失的权重。

3.2.2 输出聚合

编码器最后阶段的预测图 P4 视为最终分割图，然后通过使用 Sigmoid 函数或者 SoftMax 函数获得最终的分割输出。

4 复现细节

4.1 与已有开源代码对比

本次复现工作是基于原作者发布在 GitHub 上开源代码进行的，开源代码链接：https://github.com/SLDGroup/EMCAD在已有的源码的基础上，我主要做了以下工作：

（1）复现原文提出的 EMCAD 解码器网络结构；

（2）优化代码结构，去除冗余代码，使代码可读性更高；

（3）由于原作者只开源了 Synapse 多器官分割的代码，我根据此部分的实现，完成了对 ACDC 心脏器官分割的实验工作；

（4）原文中提到经过消融实验发现，多尺度内核 [1,3,5] 在实验效果上最好，因此我次部分进行了相关的验证。

4.2 实验环境搭建

本次实验本地代码是基于 windows10 系统，依赖于Pycharm编译器，训练依赖于服务器显卡，Python 环境如下：

Python==3.8

PyTorch==2.4.1

torchvision==0.19.1

torchaudio==2.4.1

mmcv-full==1.7.2

环境内所需其他库通过运行命令：`pip install -r requirements.txt`完成安装。

4.3 数据集合和预训练模型准备

Synapse 多器官数据集准备：先在Synapse 官方网站上注册个人账号，然后下载预处理后的数据，并保存在“./data/synapse/”文件夹中。

ACDC 数据集：从Google Drive of MT-UNet 下载预处理后的 ACDC 数据集，然后移动到“./data/ACDC”文件夹中。

4.3.1 训练模型准备

预训练模型从Google Drive/PVT GitHub下载预训练模型 PVTv2，然后将其放入“./pre-trained_pth/pvt”文件夹中。

4.4 创新点

在本文中的损失函数定义中，对每一层输出的图像（P1、P2、P3、P4）赋予的权重都为 1，表示每层分割结果对最终分割结果的影响都相同。但是在图像分割任务中，尤其是在层级（级联）结构中，为每一层分配不同的权重是一种常见的做法，以强调不同层次的特征或调整模型的学习过程，如 LC Chen [4] 讲述通过多尺度特征融合来提高分割性能，这涉及到需要对不同尺度特征的权重分配。因此基于以上，我对论文的损失函数进行了一定的修改，借鉴了 Lin [14] 论文中的思想，使其实现动态调整损失函数权重，达到更合理的逻辑设计。

5 实验结果分析

5.1 实验设置

本次工作复现是使用了 ImageNet 预训练模型的 PVTv2-b2 作为编码器。在 MECAD 编码器中的 MSDC 中，通过使用消融实验最终选择了多尺度内核 [1,3,5]。本文工作在所有实验中都是使用的深度卷积的并行排列，使用 AdamW [16] 优化器训练模型，学习率和权重衰减为 $1e-4$。其中，对于 Synapse 和 ACDC 数据集，我们将图像大小调整为 224×224，并进行随机旋转和翻转增强，优化交叉熵（0.3）和 DICE（0.7）损失组合。Synapse 多器官模型（300 epochs , batch size 6），ACDC 心脏器官模型（400 epochs , batch size 12 ），最终根据 DICE 分数保存最佳模型。

5.2 实验结果及分析

5.2.1 Synapse 腹部多器官分割

本次复现实验主要是在 Synapse 多器官数据集上进行的，据表1所示，PVT-EMCAD-B2 在腹部器官分割方面表现出色，取得了 83.63% 的最高平均 DICE 分数（复现结果达到了 83.44%），超过了所有基于 SOTA CNN 和 Transformer 的方法。他的 DICE 分数比 PVT-CASCADE 高出了 2.7%，HD95 比 PVT-CASCADE 低 4.55，这表明他在器官边界定位方面更优。

表 1. 在 Synapse 多器官数据集进行腹部器官分割结果

Architectures	Average		
	DICE↑	HD95↓	mIoU↑
UNet	70.11	44.69	59.39
AttnUNet	71.70	34.47	61.38
R50+UNet	74.68	36.87	-
R50+UNet	75.57	36.97	-
SSFormer	78.01	25.72	67.23
PolypPVT	78.08	25.61	67.43
TransUNet	77.61	26.9	67.32
SwinUNet	77.58	27.32	66.88
MT-UNet	78.59	26.59	-
MISSFormer	81.98	18.20	-
PVT-CASCADE	81.06	20.23	70.88
TransCASCADE	82.68	17.34	73.48
PVT-EMCAD-B2	**83.63**	**15.68**	**74.65**
Reproduced results	83.44	15.68	74.65

5.2.2 ACDC 心脏器官分割

表2显示了 PVT-EMCAD-B2 以及其他 SOTA 方法在 ACDC 数据集的核磁共振图像上进行心脏器官分割的 DICE 分数与 #Params 和 #FLOPs 的散点图。由表格数据很明显观察到，PVT-EMCAD-B2 获得了 92.12% 的最高平均 DICE 分数，比 Cascaded MERIT 提高了约 0.27%，但是本工作的网络计算成本要低很多。

表 2. ACDC 心脏数据集分割结果

Methods	Avg.DICE	RV	Myo	LV
R50+UNet	87.55	87.10	80.63	94.92
R50+AttnUNet	86.75	87.58	79.20	93.47
ViT-CUP	81.45	81.46	70.71	92.18
R50+ViT+CUP	87.57	86.07	81.88	94.75
MT-UNet	90.43	86.64	89.04	95.62
MISSFormer	90.86	89.55	88.04	94.99
PVT-CASCADE	91.46	89.97	88.9	95.50
TransCASCADE	91.63	90.25	89.14	95.50
Cascaded MERIT	91.85	90.23	89.53	95.80
PVT-MECAD-B2	**92.12**	**90.65**	**89.68**	**96.02**

5.2.3 二值医学图像分割结果分析

如图3和图4所示是二值医学图像在不同方法上的分割结果，可以看到 PVT-EMCAD-B2 取得了最高的平均 DICE 分数，为 91.10%，但仅需要 26.76M 的 Params 和 5.6G 的 FLOPs，这表明 EMCAD 解码器中的多尺度深度卷积与 Transformer 编码器相结合，可以减少参数、提高计算性能的同时，保持或提高分割的准确率。

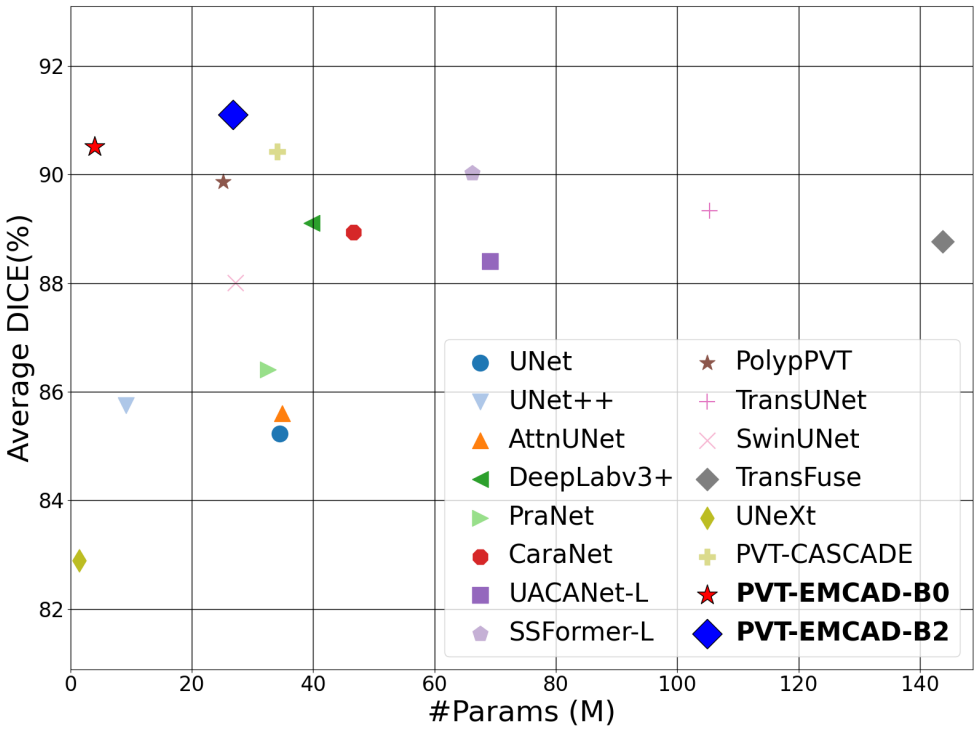

图 3. 不同方法的平均 DICE 分数与 #Params 散点图

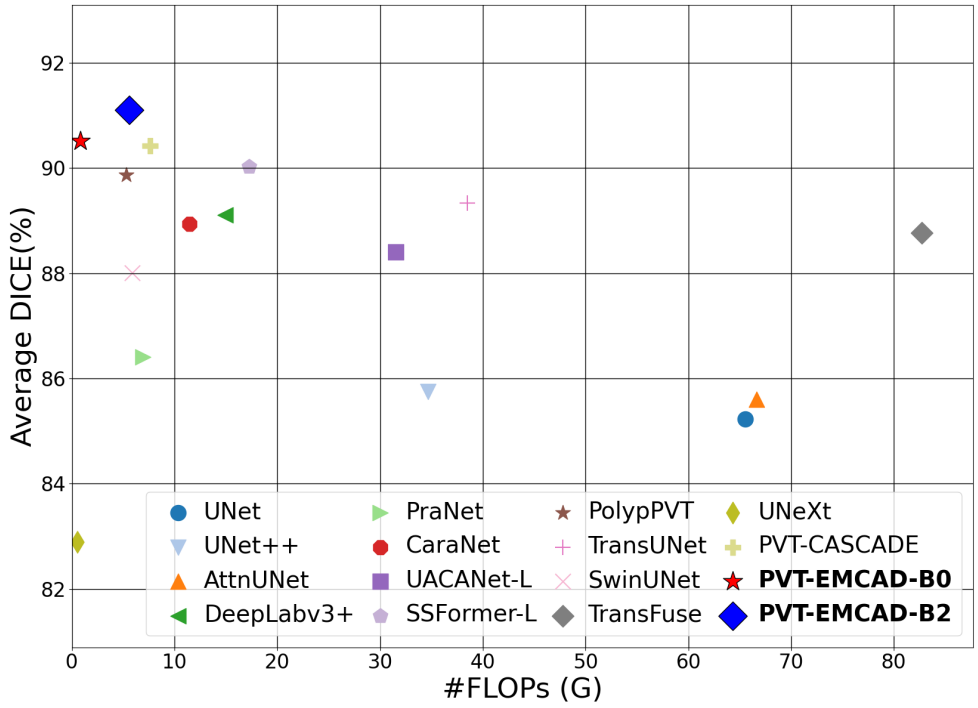

图 4. 不同方法的平均 DICE 分数与 #FLOPs 散点图

6 总结与展望

本次工作聚焦于复现 Rahman 提出的 EMCAD，这是一种高效的多尺度卷积注意力解码器，专为医学图像分割中的多级特征聚合和细化而设计的。EMCAD 采用了多尺度深度卷积块，这是捕捉特征图中不同尺度信息的关键，也是医疗图像分割精度的关键因素。采用深度卷积而不是标准的 3×3 卷积块，这种设计选择使 EMCAD 具有显著的效率提升。实验表明，EMCAD 在 DICE 分数上超过了 CASCADE 解码器，参数减少了 79.4%，FLOPs 减少了 80.3%。同时，EMCAD 的设计使其能够适应不同的编码器，并在多种分割任务中表现出色，这种灵活性也意味着它可以在不同的应用场景中高效地工作，包括那些计算资源受限的环境。实验证明 EMCAD 与小型编码器的兼容性使其在保持高性能的同时，也非常适合医疗点的应用。

尽管本文提出的 EMCAD 解码器在医学图像分割任务中表现出色，但我觉得在以下几个方面仍有一些潜在的改进方向。

（1）本文主要关注的是二维医学图像分割，而许多医学成像技术（如 CT 和 MRI）生成的是三维图像。将 EMCAD 扩展到三维图像分割的任务中，处理三维数据的复杂性和计算需求会更高，但也将进一步扩展其应用范围。

（2）在实际的应用中，医学图像可能会受到噪声、模糊或不完整数据的影响，可以进一步研究如何提高 EMCAD 对这些不利因素的鲁棒性，以确保在各种条件下都能稳定地进行图像分割。

（3）在不同的医学成像模态（如 X 射线、超声、MRI 等）具有不同的图像特征和噪声特征。虽然 EMCAD 在多个数据集上表现良好，但可以进一步研究如何使其更好地适应不同模态的图像，以提高在特定模态下的性能。

（4）提高模型的可解释性，使医生和研究人员能够更好地理解模型的决策过程和依据，这对于医学图像分析中的信任和应用至关重要。

参考文献

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla SegNet. A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 5, 2015.

[2] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[3] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[4] Liang-Chieh Chen. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[7] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*, 2021.

[8] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 263–273. Springer, 2020.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Andrew G Howard. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[12] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1055–1059. IEEE, 2020.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[14] T Lin. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017.

[15] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[16] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[17] Ange Lou, Shuyue Guan, and Murray Loew. Dc-unet: rethinking the u-net architecture with dual channel efficient cnn for medical image segmentation. In *Medical Imaging 2021: Image Processing*, volume 11596, pages 758–768. SPIE, 2021.

[18] Md Mostafijur Rahman and Radu Marculescu. Multi-scale hierarchical vision transformer with cascaded attention decoding for medical image segmentation. *arXiv e-prints*, pages arXiv–2303, 2023.

[19] Md Mostafijur Rahman and Radu Marculescu. Medical image segmentation via cascaded attention decoding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6222–6231, 2023.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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*, pages 234–241. Springer, 2015.

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[22] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[23] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[24] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pages 459–479. Springer, 2022.

[25] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021.

[26] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022.

[27] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[28] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[29] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018.