UTNet: A Hybrid Transformer Architecture for Medical Image Segmentation

Yunhe Gao, Mu Zhou, Dimitris Metaxas

摘要

Transformer 架构已经在许多自然语言处理任务中取得成功。然而，它在医学视觉中的应用在很大程度上仍未得到探索。在这项研究中，本文提出了 UTNet，这是一种简单而强大的混合 Transformer 架构，它将自注意力集成到卷积神经网络中，以增强医学图像分割。UTNet 在编码器和解码器中应用自注意力模块，以最小的开销捕获不同规模的远程依赖。为此，我们提出了一种有效的自注意力机制以及相对位置编码，将自注意力操作的复杂性从 $O(n^2)$ 显著降低到近似 $O(n)$。还提出了一种新的自注意力解码器，以从编码器中跳过的连接中恢复细粒度的细节。本文所提的方法解决了 Transformer 需要大量数据来学习视觉归纳偏差的困境。同时混合层设计允许在不需要预训练的情况下将 Transformer 初始化为卷积网络。我们在多标签、多供应商心脏磁共振成像队列图像中评估了 UTNet。大量实验数据显示出 UTNet 优越的分割性能和对最先进方法的鲁棒性，有望在其他医学图像分割中得到很好的推广。

关键词：医学图像分割；UTNet；Transformer；Self-Attention

1 引言

卷积网络以其卓越的特征表示能力革命性地改变了计算机视觉领域。目前，卷积 encoder-decoder 架构在位置敏感任务方面取得了实质性进展，如语义分割[1-5]。借助卷积操作可以从相邻像素中收集局部信息来捕获纹理特征，为了能全局地聚合局部滤波器响应，这些模型通过堆叠多个卷积层和下采样操作扩大感受野。尽管卷积操作有这些有点，但这种模式有两个固有的局限性。首先，卷积只从邻域像素收集信息，缺乏显式捕获长距离（全局）依赖关系的能力[6-8]。二是卷积核的尺寸和大小往往是固定的，无法根据输入内容进行调整[9]。

Transformer 已经被广泛用于 NLP 领域[10]，其 self-attention 机制可以有效的捕获长程依赖关系。自注意力机制会计算像素之间的交互关系，聚合之后作为输出，可以根据输入内容动态聚合相关特征，也可以有效的捕获长程关联。已经有研究初步表明，self-attention 在分割[11-12]、检测[13]、图像复原[14]等领域均十分有效。

尽管将 Transformer 应用于视觉任务十分具有前景，但是具体的应用和实现还存在一些挑战。首先，自我注意机制具有 $O(n^2)$ 时间和空间复杂度，n 是序列长度，因此会有大量的训练和推理开销。一些前人的工作尝试简化 self-attention 的时间复杂度[15-16]，但仍远未达到完美，仍有很大的优化空间。受限于 self-attention 的复杂度，在 ViT 中只能将图像切分为 16×16 大小图像 patch 块作为输入序列[17-18]；或者采用 CNN 提取到的已经降为过的 feature map 作为输入[11,19]。但是对于医学图像分割任务，细节位置信息往往对精确分割更为重要，因为大多数误分割区域都位于边缘；其次，由于 Transformer 没有引入任何归纳偏置，导致其在小规模数据集上表现不，Transformer 一般都需要现在大型数据集 (如

JFT-300M[17]) 上进行预训练再进行迁移。但是即使在 ImageNet 预训练过，Transformer 的表现仍然稍逊于 ResNet[20-21]，更遑论医学图像数据集这种数据集规模更小的情况。

为此，本论文提出了混合 Transformer 的 U 型网络-UTNet，充分将 self-attention 和卷积结合在一起，利用卷积层提取足够的局部特征，同时避免 Transformer 对大规模数据集的预训练需求；同时借助 self-attention 捕获长程信息、全局特征。UTNet 网络设计遵循标准的 UNet 结构，但是将每层卷积模块中的最后一个卷积操作替换为 Transformer 结构。为了进一步提升分割效果，本文还在较高分辨率的 feature map 使用 self-attention 来提取细节之间的长程依赖，并且将 self-attention 的计算时间和空间复杂度从 $O(n^2)$ 降低至 $O(n)$；此外本文还引入了相对位置编码来学习医学图像中内容-位置的关联。通过在多器官分割、心室分割任务中的实验表明，UTNet 均表现出了优异的分割性能和鲁棒性；UTNet 这种设计也有望推广到其他医学图像分割任务。

2 相关工作

2.1 传统医学图像分割

医学图像分割是计算机视觉领域重要的研究方向，为了获取准确的分割结果，细节信息和全局信息都很重要。传统的 UNet 利用它特殊的对称结构在高分辨率图像中获取局部特征，低分辨率图像中捕捉全局特征，实现端到端的分割[1]。UNet 在编码器和-解码器结构中，结合上下采样和跳跃连接，融合多尺度特征信息，为分割模型提供了粗细特征图的同时还能加速模型收敛，对于处理医学图像分割任务极其有效。不仅如此，作者提出的 UNet 不包含全连接层，而是使用参数量少的全卷积层代替。U-Net++ 为了能够减小编码器和解码器特征图之间的差异，进一步改进了 U 型网络[22]。U-Net++ 在跳跃连接上加上了若干卷积层，并在各卷积层之间使用密集连接 (Dense Connection)[23]，以减小两边网络特征表达的差异。此外，作者把编码器每阶段产生的特征图通过上采样到原图大小，然后和标签计算损失，监督特征融合操作。ResUNet (Residual and U-Net)[24]把 UNet 所提出模型的骨干网络的卷积部分用残差网络 (Residual Network，ResNet)[21]代替，在此的基础上，ResUNet++[25]在 ResUNet 编码器中的每个残差块之后添加压缩提取模块 (Squeeze and Extraction Block, SE Block)[26]不仅把编码器中不同尺度的特征图传递给解码器，还传递了通道注意力权重。利用注意力权重过滤掉解码器特征图的多余信息，再将其输入到后面的网络中。相关实验表明，这种融合两边网络特征的方式比起一次性串联更加有效。

2.2 基于 Transfomer 的 U 型分割网络

在 Transformer 被应用到医学图像分割领域之前，基于 CNN 或全卷积网络的分割模型在各大图像分割下游任务中展现着出色的表现。但随着 Transformer 在 NLP 任务中大放光彩，ViT[17]将 Transformer 应用到图像分类任务中并取得成功之后，Chen 等[27]提出 TransUNet(Transformers and U-Net)。该模型的出现开启 Transformer 在医学图像分割领域中的应用。由于 Transformer 在大规模数据集上才能更好的发挥其优势，而大多数医学图像数据属于小规模数据集，因此，研究进一步改进 Transformer 模块使其适用于医学图像处理便成了热门的研究方向之一。其中，最为有效的方法之一就是把 Transformer 同 U 型网络结合，利用 U 型网络尽可能减小计算量的同时也能有效捕捉重要信息的特点，充分挖掘 Transformer 和 U 型网络的潜力。

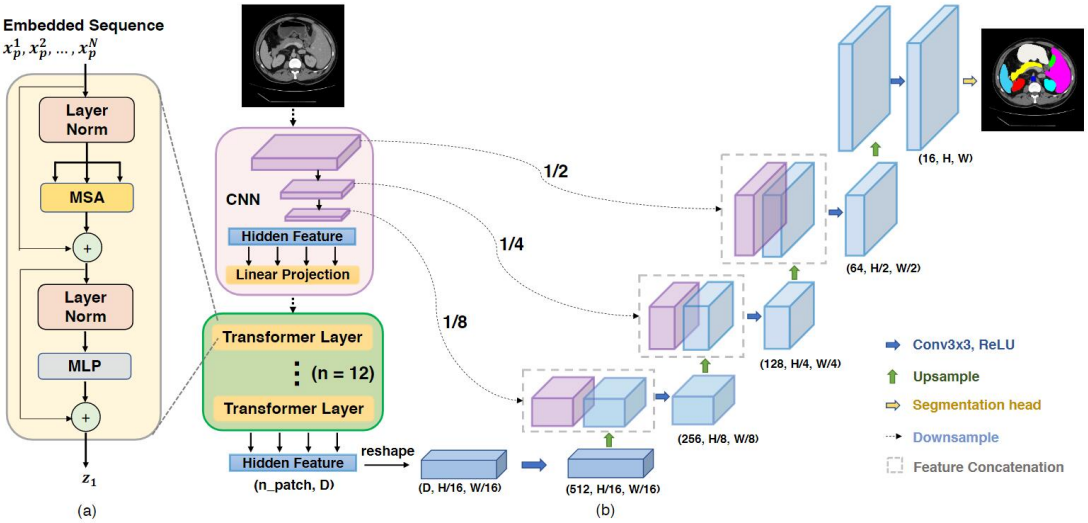

图 1: TransUnet 网络结构图

TransUnet 是首个将 Transformer 应用到医学图像分割领域的 U 型网络，如 1 所示。该模型直接将编码器中下采样之后的图像序列化，然后套用到 NLP 中最原始的 Transformer 模块中进行训练。利用 Transformer 在低分辨率图像中获取长距离依赖的优势和对称的编码器-解码器结构，提升了模型自动分割的性能。但由于 TransUNet 直接使用了 NLP 的 Transformer 模型，其序列中的图像块大小固定，并且由于注意力计算量大，所以 TransUNet 的分割效率还有待进一步的提升。

nnFormer(not another transFormer)[28]在网络中交替使用 Transformer 和 CNN，并提取每一尺度的特征信息进行多尺度监督学习，保证多尺度的特征表达尽可能准确。但引入多个 Transformer 会大大增加计算负载，于是作者将 Transformer 提前在 Imagenet 中预训练之后，固定注意力模块和多层感知机(Multi-Layer Perceptron, MLP)层参数，其他部分根据目标任务进行新的学习。另外，受 Swin Transformer 启发，作者用三维窗口替换原来的二维窗口，在窗口内进行自注意力计算，相较于原始的三维多头注意力机制，计算量减少了 90% 以上。为了避免三维窗口和三维图像不匹配而导致计算时填充冗余信息，三维窗口大小根据三维图像专门设定。不仅如此，作者提出用连续的、小的卷积层比 ViT 中直接用单个的、大的卷积层学到的嵌入层有着更丰富的位置信息，还有助于降低模型复杂度。

3 本文方法

3.1 本文方法概述

本文使用一种融合 Transfome 的 U 型网络 UTNet，充分融合卷积和自注意力的优点，获取长距离信息的同时降低训练的数据需求。为了降低计算复杂度，作者提出一种有效的自注意力机制，将时间复杂度和空间复杂度从 $O(n^2)$ 降低到接近 $O(n)$，并使用一种相对位置编码学习高度结构化的图像数据之间的位置关联。

3.2 有效的自注意力机制

图像是高度结构化的数据，在局部边界区域的大部分像素都具有相似的特征，所以在计算所有像素点的成对注意力是存在大量冗余的，因此计算也十分低效。已经有研究通过理论分析证明，对于一些较长序列，其 self-attention 的都是低秩矩阵[29]，表明其绝大多数信息都集中在较大的其奇异值上。基于此启发，本文提出了一种更高效的 self-attention 计算方法，如图 2所示。

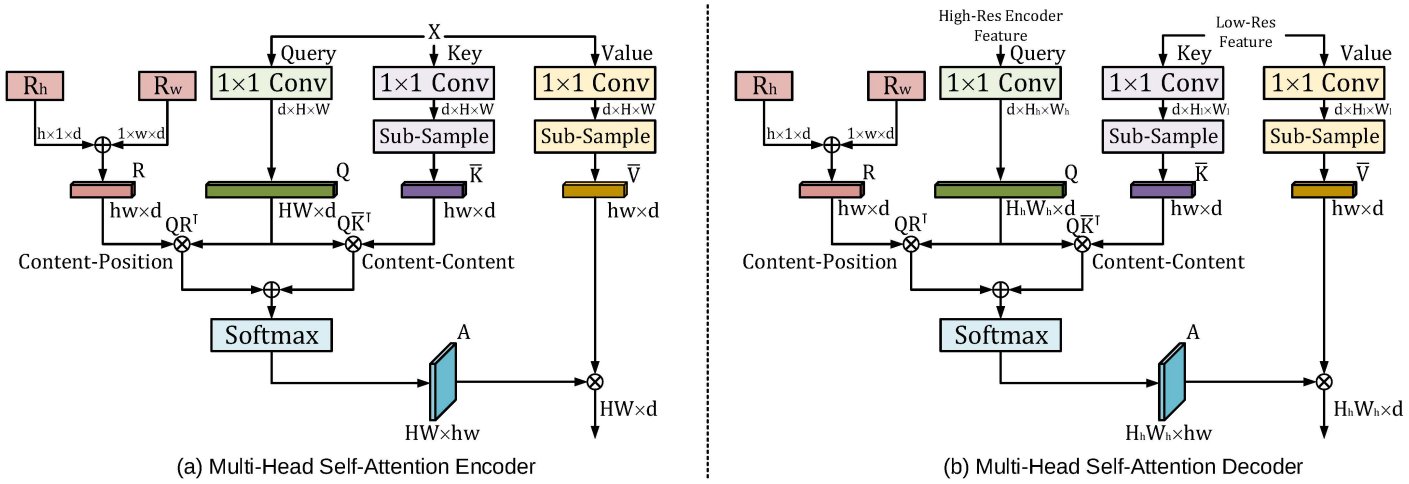

图 2: 有效的过头注意力机制 (MHSA). **(a)** 编码器中的多头自注意力机制。**(b)** 解码器中的多头自注意力机制。

本文的主要的想法是分别将高维的 key 和 value：$\mathbf{K}, \mathbf{V} \in \mathcal{R}^{n \times d}$ 映射到低维的 embedding: $\overline{\mathbf{K}}, \overline{\mathbf{V}} \in \mathcal{R}^{k \times d}$, 其中 h 和 w 为下采样后特征图的大小，特征图投影后的维度 $k = hw \ll n$。并且计算时使用的是 K,V 的低维近似，这样可以减少计算量，将时间复杂度降到 $O(nkd)$。

$$\text{Attention}(\mathbf{Q}, \overline{\mathbf{K}}, \overline{\mathbf{V}}) = \underbrace{\text{softmax}(\frac{\mathbf{Q}\overline{\mathbf{K}}^{\intercal}}{\sqrt{d}})}_{\overline{P}:n \times k} \underbrace{\overline{\mathbf{V}}}_{k \times d} \tag{1}$$

值得注意的是：低维近似可以是任何降采样操作，如均值池化、最大池化、带步长的卷积等，本文则是使用 1x1 卷积接双线性插值来对特征图进行降采样。图 2 分别展示了在 Transformer 编码器和解码器中使用的多头自注意力模块（MHSA），他们流程相似，但是 decoder 中输入有两个，一个是从编码器通过 skip connection 连接过来的特征图，一个则是解码器前一层的低分辨率的特征图。

3.3 相对位置编码

标准的自注意力模块完全丢弃了位置信息，使其是扰动等变的[30]，从而模型无法对结构信息进行建模。最初使用的正弦位置 embedding 方式也并不具有像 CNN 的平移不变性性质[31]。本文使用 2D 的相对位置编码方式，在进行 softmax 之前为每一个像素点额外的加入高度和宽度信息：

$$l_{i,j} = \frac{q_i^{\intercal}}{\sqrt{d}}(k_j + r^W_{j_x - i_x} + r^H_{j_y - i_y}) \tag{2}$$

其中 q_i 是像素 i 的 query 向量, k_j 是像素 j 的 key 向量, $r^W_{j_x - i_x}$ 和 $r^H_{j_y - i_y}$ 是相对宽度 $j_x - i_x$ 和相对宽度 $j_y - i_y$ 的可学习的 embedding。与 efficient self-attention 计算相似，相对宽度和高度也是取低维映射后的结果。将位置编码加上的自注意力机制公式 3所示：

$$\text{Attention}(\mathbf{Q}, \overline{\mathbf{K}}, \overline{\mathbf{V}}) = \underbrace{\text{softmax}(\frac{\mathbf{Q}\overline{\mathbf{K}}^{\intercal} + \mathbf{S}^{rel}_H + \mathbf{S}^{rel}_W}{\sqrt{d}})}_{\overline{P}:n \times k} \underbrace{\overline{\mathbf{V}}}_{k \times d} \tag{3}$$

其中 $\mathbf{S}^{rel}_H, \mathbf{S}^{rel}_W \in \mathcal{R}^{HW \times hw}$ 是相对位置编码矩阵，满足 $\mathbf{S}^{rel}_H[i,j] = q_i^{\intercal} r^H_{j_y - i_y}, \mathbf{S}^{rel}_W[i,j] = q_i^{\intercal} r^W_{j_x - i_x}$。

3.4 整体网络架构

UTNet 的整体网络结构如图 3所示。我们尝试将卷积与 self-attention 的优势结合在一起，这样一方面借助卷积可以学习归纳偏置，同时避免了在大规模数据集上预训练 Transformer 的需求；另一方面利用 Transformer 对捕获全局特征。通过本文提出的有效的自注意力机制和相对位置编码，使得我们可以

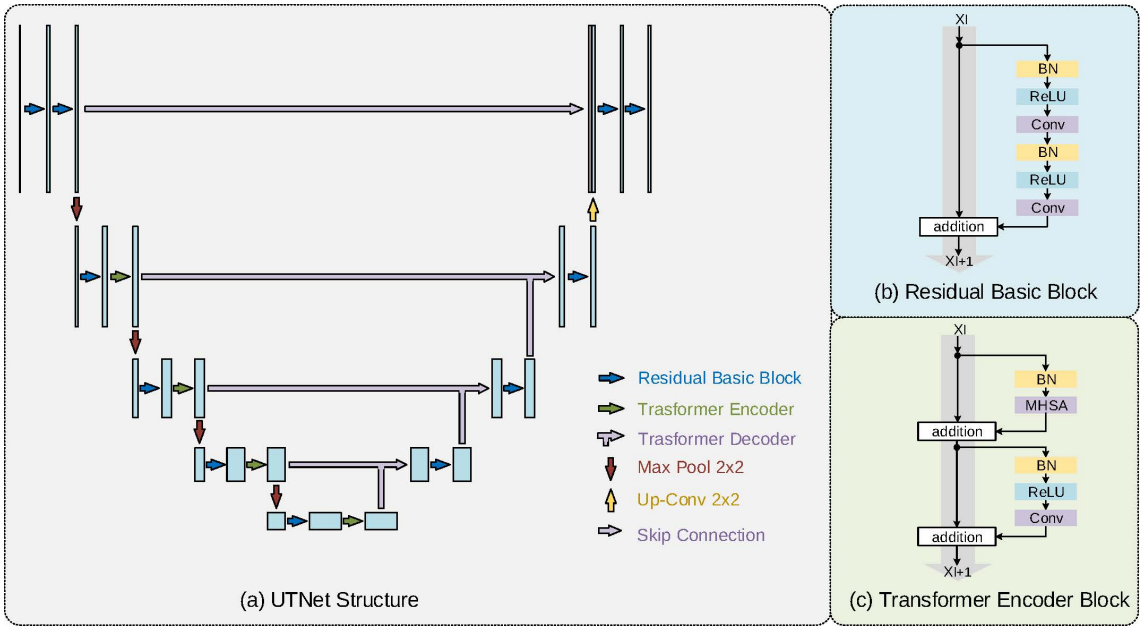

图 3: **(a)** UTNet 网络总体结构。有效地自注意力机制和相对位置编码让我们能够应用 Transfomer 结构整合编码器和解码器中不同大小的全局信息。**(b)** 基础的残差连接. **(c)** Transformer 编码器块结构.

借助 Transformer 有效的聚集不同尺度的全局上下文信息。由于误分割经常发生在 ROI 区域的边界，因此高分辨率的上下文信息就对精确分割至关重要。本文设计的焦点就在于如何使用 self-attention 模块高效地计算大尺寸的 feature map。我们并不是简单的在 CNN 提取的 feature map 上计算 self-attention，而是在编码器-解码器每一级别上使用 Transformer 来收集不同尺度的长距离依赖关系。特别的是原始输入并没有使用 Transformer 处理，因为在较浅层次使用 Transformer 不仅没什么额外的作用还会增加计算成本。一种可能的解释是是浅层 feature map 更关注细节纹理，而此时获取地全局信息可能并不重要。图 3 中 (b), (c) 是残差连接块和 Transformer 块地具体细节。

4 复现细节

5 复现细节

5.1 与已有开源代码对比

本次复现是基于作者原有代码[1]进行改进，受 PHTrans[32]工作启发，本次改进主要将原来交替串行的 Transformer 和 CNN 网络结构改成为并行的特征提取模块。交替的 Transformer 和 CNN 架构不能利用连续的模型全局或局部的表示，将模型全局信息同局部信息直接糅合而不能充分利用 Transformer 和 CNN 的应用潜力。而改进后的网络结构能够同时独立地利用 Transformer 结构和 CNN 结构获取局部和全局信息，并在不同阶段融合全局和局部信息，充分发挥二者的潜力。

5.2 实验环境设置

本实验将 UTNet 在多器官分割、心室分割任务上进行了测验，包括左心室 (LV)、右心室 (RV) 和左心室心肌 (MYO) 的分割[33]。在训练集中，我们有来自两个不同 MRI 供应商的 150 张带注释的图像 (每个供应商的 75 张图像)，包括 A：西门子；B：飞利浦。在测试集中，我们有来自 4 个不同 MRI 供应商的 200 个图像 (每个供应商 50 个图像)，包括 A：西门子；B：飞利浦；C：GE；D：佳能，其中供应商 C 和 D 的图像数据没有混入训练集中。由于来自不同供应商的 MRI 扫描图像在外观上有明显的

[1]https://github.com/yhygao/UTNet

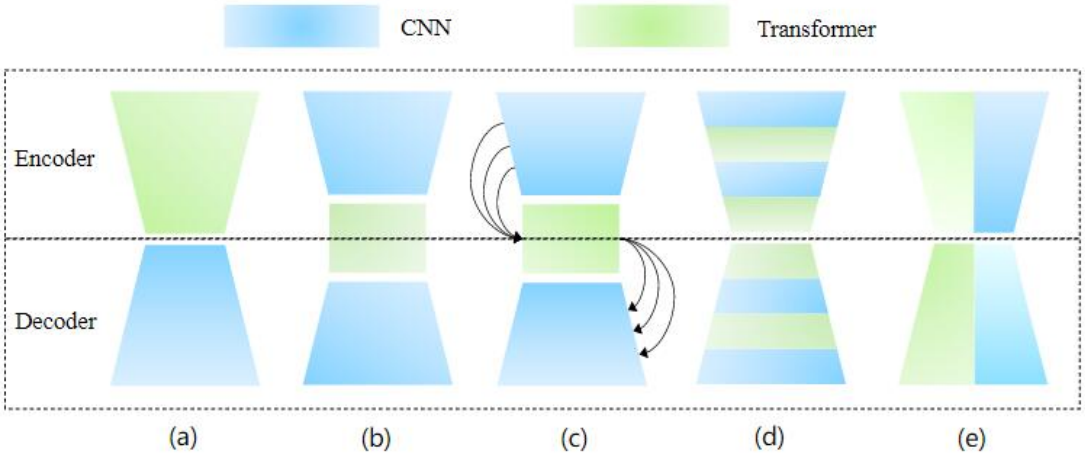

图 4: 不同的 Transformer 和 CNN 杂合的架构，**(a)**Tranformer 仅作为编码器；**(b)** Transformer 在中间层作为编码器-解码器中间层；**(c)**Transformer 作为桥梁连接编码器和解码器；**(d)**Transformer 和 CNN 在编码器-解码器中交替使用；**e**Transformer 和 CNN 并行结构。

差异，我们能够测量模型的鲁棒性。更具体地说，我们进行了两个实验将 Dice 分数和 Hausdorff 距离作为评价指标来验证 UTNet 的性能和鲁棒性。我们测量了训练和测试数据都来自同一供应商 A 的初步结果。其次，我们进一步衡量了模型的跨供应商间的鲁棒性。

5.3 复现细节

对于数据预处理，我们将平面间距重新采样为 1.2×1.2mm，同时保持沿 z 轴的间距不变。我们将所有的模型从零开始训练了 150 个 epochs。训练时我们使用基础学习率为 0.05 的指数学习率衰减。我们在单个 GPU 上使用小批量随机梯度下降算法 SGD 训练，batch size 为 16，下采样大小 reduce size 设置为 8，动量和重量衰减设置为 0.9 和 1e−4。在模型训练过程中动态使用数据增强，包括随机旋转、缩放、平移、可加性噪声和伽马变换，最后所有图像都被随机裁剪为 256×256。我们使用 Dice 损失和交叉熵损失的组合来训练所有网络。

5.4 创新点

仅采用 CNN 架构较难捕获长程全局信息，而单纯的 Transformer 架构由于参数量和计算复杂度较高需要大规模的预训练。将两种架构混合到一个模型中能够同时利用各自的优点，然而常见串行的交替的架构不能充分利用二者的优越性。本次实验通过使用 Transformer 和 CNN 融合的并行的架构，让模型获取的全局信息和局部信息能在模型的不同阶段得到不同层级更为准确的特征表示，充分将二者特有架构的潜力发挥出来。

6 实验结果分析

表 1 展示了不同网络结构的泛化性能，即测试和训练使用不同的数据集，UTNet 仍然展现了优异的性能。这一结果可以归因于 UTNet 在不同层次 feature map 上均使用了 self-attention 来聚合全局信息、内容-位置之间的关联信息，这样赋予了 UTNet 更好的关注全局上下文的能力，而不仅仅是关注局部纹理。图 5也显示出 UTNet 最一致的分割边界（特别是心脏核磁共振扫描图像中的 RV 和 MYO 区域），表明 UTNet 利用 Transformer 后的有效性和鲁棒性。

改进后的结果如表 2所示，可以改进后不同的分割目标的 Dice 分数都有所提升，同时对不同分割目标的 Hausdorff 距离进行评估也能产生类似的结果，表明我们通过并行化 Transformer 和 CNN 架构的改进是有效的。

表 1: 不同模型下的 Dice 分数。所有的模型均只使用 A，B 两个供应商的数据作为训练集，A、B、C 和 D 作为测试集。括号中的向下的箭头和值表示相对于 A，B 平均分数的下降的值。

	ResUNet				CBAM				UTNet			
Vendor	A	B	C	D	A	B	C	D	A	B	C	D
LV	92.5	90.1	88.7 (↓2.6)	87.2 (↓4.1)	**93.3**	91.0	89.4 (↓2.8)	88.8 (↓3.4)	93.1	**91.4**	**89.8 (↓2.5)**	**90.5 (↓1.8)**
MYO	83.6	85.3	82.8 (↓1.7)	80.2 (↓4.3)	**83.9**	85.8	82.6 (↓2.3)	80.8 (↓4.1)	83.7	**85.9**	**83.7 (↓1.1)**	**82.6 (↓2.2)**
RV	87.4	87.5	85.9 (↓1.6)	85.3 (↓2.2)	88.4	88.4	85.3 (↓3.1)	86.4 (↓2.0)	**89.4**	**88.8**	**86.3 (↓2.8)**	**87.3 (↓1.8)**
AVG	87.9	87.6	85.7 (↓2.0)	84.2 (↓3.5)	88.5	88.4	85.5 (↓2.7)	85.3 (↓3.2)	**88.7**	**88.7**	**86.6 (↓2.1)**	**86.2 (↓2.5)**

表 2: 改进前后模型精度对比

	改进前				改进后			
Vendor	A	B	C	D	A	B	C	D
LV	93.05	**91.46(↑ 0.24)**	**88.94 (↑ 0.09)**	88.45	**93.35(↑ 0.3)**	91.22	88.85	**88.75(↑0.3)**
MYO	**83.88(↑0.07)**	85.84	82.4	80.71	83.81	**85.91(↑0.07)**	**82.89(↑0.49)**	**81.59(↑0.88)**
RV	88.48	88.31	84.11	85.63	**88.58(↑0.1)**	**88.53(↑0.22)**	**85.19(↑0.18)**	**85.81(↑0.18)**
AVG	88.37	88.46	85.12	84.99	**88.59(↑0.22)**	**88.62(↑0.16)**	**85.59(↑0.47)**	**85.32(↑0.33)**

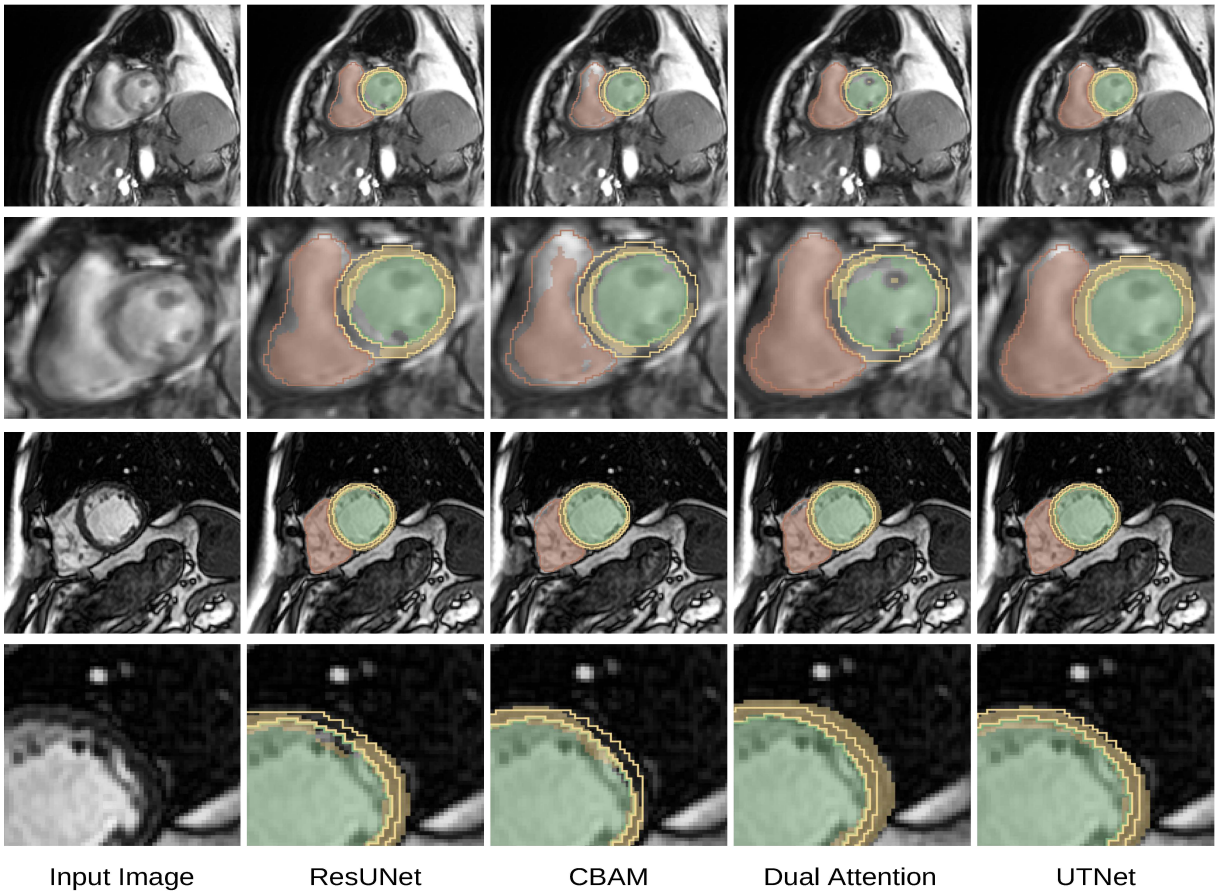

图 5: 非训练集中供应商 C 和 D 中的图片的分割效果图。绿色表示左心室，黄色表示心肌细胞，红色表示右心室；供应商 C 的图片由于运动伪影而模糊，而供应商 D 的图片噪声大，边界对比度低。仅 UTNet 提供一致的分界，这证明了它的鲁棒性。

7 总结与展望

由于作者提供的实验代码有些混乱，在其代码上的改进比较困难。Transformer 融合 CNN 结构的方向仍然大有可为，详实的实验数据可以提供更为坚实的证据，模型中的双线性插值的下采样操作有些过于草率，经过仔细设计的下采样操作如空洞卷积或小波重建可以保留更细节的全局或局部信息，

另外借鉴 Swin Transformer[34]将图像成 patch 之后再做投影，而不是直接对单个像素直接投影至 32 维是否会更为有效也值得思考，同时像 PSPNet[6]中提出多尺度特征融合的一样，将不同下采样大小的特征图做融合也值得探索。

参考文献

[1] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. 2015: 234-241.

[2] ISENSEE F, JAEGER P F, KOHL S A, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation[J]. Nature methods, 2021, 18(2): 203-211.

[3] WANG S, ZHOU M, LIU Z, et al. Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation[J]. Medical image analysis, 2017, 40: 172-183.

[4] TAJBAKHSH N, JEYASEELAN L, LI Q, et al. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation[J]. Medical Image Analysis, 2020, 63: 101693.

[5] GAO Y, LIU C, ZHAO L. Multi-resolution path cnn with deep supervision for intervertebral disc localization and segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. 2019: 309-317.

[6] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2881-2890.

[7] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[J]. arXiv preprint arXiv:1511.07122, 2015.

[8] GAO Y, HUANG R, YANG Y, et al. FocusNetv2: Imbalanced large and small organ segmentation with adversarial shape constraint for head and neck CT images[J]. Medical Image Analysis, 2021, 67: 101831.

[9] SCHLEMPER J, OKTAY O, SCHAAP M, et al. Attention gated networks: Learning to leverage salient regions in medical images[J]. Medical image analysis, 2019, 53: 197-207.

[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is All you Need[C]//NIPS. 2017.

[11] FU J, LIU J, TIAN H, et al. Dual attention network for scene segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3146-3154.

[12] SINHA A, DOLZ J. Multi-scale self-guided attention for medical image segmentation[J]. IEEE journal of biomedical and health informatics, 2020.

[13] YI J, WU P, JIANG M, et al. Attentive neural cell instance segmentation[J]. Medical Image Analysis, 2019, 55: 228-240. DOI: https://doi.org/10.1016/j.media.2019.05.004.

[14] HUANG Q, YANG D, WU P, et al. MRI reconstruction via cascaded channel-wise attention network[C]//2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). 2019: 1622-1626.

[15] HUANG Z, WANG X, HUANG L, et al. Ccnet: Criss-cross attention for semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 603-612.

[16] ZHU Z, XU M, BAI S, et al. Asymmetric non-local neural networks for semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 593-602.

[17] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[18] ZHENG S, LU J, ZHAO H, et al. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers[J]. arXiv preprint arXiv:2012.15840, 2020.

[19] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7794-7803.

[20] KOLESNIKOV A, BEYER L, ZHAI X, et al. Big transfer (bit): General visual representation learning [J]. arXiv preprint arXiv:1912.11370, 2019, 6(2): 8.

[21] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[22] ZHOU Z, SIDDIQUEE M M R, TAJBAKHSH N, et al. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation[J]. IEEE transactions on medical imaging, 2019, 39(6): 1856-1867.

[23] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

[24] DIAKOGIANNIS F I, WALDNER F, CACCETTA P, et al. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 162: 94-114.

[25] JHA D, SMEDSRUD P H, RIEGLER M A, et al. Resunet++: An advanced architecture for medical image segmentation[C]//2019 IEEE International Symposium on Multimedia (ISM). 2019: 225-2255.

[26] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[27] CHEN J, LU Y, YU Q, et al. Transunet: Transformers make strong encoders for medical image segmentation[J]. arXiv preprint arXiv:2102.04306, 2021.

[28] ZHOU H Y, GUO J, ZHANG Y, et al. nnformer: Interleaved transformer for volumetric segmentation [J]. arXiv preprint arXiv:2109.03201, 2021.

[29] WANG S, LI B, KHABSA M, et al. Linformer: Self-attention with linear complexity[J]. arXiv preprint arXiv:2006.04768, 2020.

[30] BELLO I, ZOPH B, VASWANI A, et al. Attention augmented convolutional networks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 3286-3295.

[31] PARMAR N, VASWANI A, USZKOREIT J, et al. Image transformer[C]//International Conference on Machine Learning. 2018: 4055-4064.

[32] LIU W, TIAN T, XU W, et al. Phtrans: Parallelly aggregating global and local representations for medical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. 2022: 235-244.

[33] CAMPELLO V M, PALOMARES J F R, GUALA A, et al. Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge[Z]. 2020.

[34] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.