SwinCross：用于 PET/CT图像头颈部肿瘤分割的跨模态 Swin Transformer

Gary Y. Li、Junyu Chen、Se-In Jang、Kuang Gong 和 Quanzheng Li

摘要—西妥昔单抗联合放射治疗（RT）是无法手术的头颈癌患者的标准治疗方法。头颈部肿瘤分割是放射治疗计划制定的前提，但这一过程十分耗时。近年来，深度卷积神经网络已成为自动图像分割的事实标准。然而，由于扩大深度卷积神经网络（DCNN）视野所需的计算成本高昂，其建模长程依赖的能力仍然有限，因此，对于背景上下文跨越较长距离的目标，分割性能可能不够理想。另一方面，Transformer模型已在医学图像的多个语义分割任务中展现出捕获此类长程信息的卓越能力。受视觉 Transformer 近期成功及多模态图像分析进展的启发，我们提出了一种全新的分割模型——跨模态 Swin Transformer（SwinCross），并引入跨模态注意力（CMA）模块，以在多个分辨率下进行跨模态特征提取。为验证所提方法的有效性，我们在 HECKTOR 2021挑战赛数据集上进行了实验，并将其与 nnU-Net**[1]**（HECKTOR 2021中排名前五的方法所用的骨干网络）及其他先进的基于 Transformer 的方法（如 UNETR**[2],** 和 Swin UNETR**[3]**）进行了比较。实验结果表明，对于头颈部肿瘤分割任务，所提方法优于这些对比方法，**这得益于 CMA 模块能够捕捉**PET与CT之间更优的互补特征表示。

索引词— Transformer、网络架构、肿瘤分割、PEC/CT

I. 引言

H头颈部（H&N）癌症是全球最常见的癌症之一[4], , 约占美国所有癌症的 4%。氟代脱氧葡萄糖正电子发射断层成像和 CT 成像是头颈部癌症初始分期和随访的金标准。医学图像中的定量图像生物标志物（如影像组学）已显示出优化患者护理的巨大潜力，尤其适用于头颈部肿瘤 [5]。然而，影像组学分析依赖于一种成本高昂且容易出错的人工流程，需在三维空间中标注感兴趣体积（VOI）。因此，对 PET/CT图像中的头颈部肿瘤进行自动分割，有望支持在大规模队列中验证影像组学模型，并实现最佳的可重复性。此外，自动分割算法还可加快临床工作流程。PET 和 CT 分别侧重于代谢特征和解剖特征，在头颈部原发肿瘤分割中提供互补且协同的信息。

近年来，Transformer（一种基于自注意力机制、用于计算特征表示和全局依赖关系的神经网络）在自然语言处理和计算机视觉领域蓬勃发展 [6]。在计算机视觉领域，基于Transformer的架构取得了显著成功，并在多种任务中展现出卓越性能，包括视觉识别 [7, 8], 、目标检测 [9, 10], 、语义分割 [11, 12], 等 [8, 13-15]。视觉 Transformer在计算机视觉领域的成功，推动了其在医学影像中的应用，并在分类 [16-18] 、分割 [2, 19, 20] 和配准 [21, 22]等多种应用中展现出良好潜力。Chen等人首次提出了用于医学图像分割的 TransUNet [19] ，该方法使用 12层视觉Transformer提取瓶颈特征，采用二维UNet设计，并在中间结构中引入Transformer模块。后来

本研究得到了美国国立卫生研究院（NIH）项目（编号 NIH C06 CA059267）的资助。(通讯作者：Gary Y. LiGary Y. Li 任职于飞利浦北美研究院（地址：马萨诸塞州剑桥市，邮编 02141；电子邮箱：bettergary@gmail.com）。本研究是在其于高级医学计算与分析中心、马萨诸塞州总医院/哈佛医学院（地址：波士顿，邮编 02114）从事博士后研究期间完成的。Se-In Jang、Kuang Gong 和 Quanzheng Li 任职于高级医学计算与分析中心、马萨诸塞州总医院/哈佛医学院（地址：波士顿，邮编 02114；电子邮箱：sjang7@mgh.harvard.edu; kgong@mgh.harvard.edu; quanzheng@mgh.harvard.edu）。Junyu Chen 任职于约翰斯·霍普金斯大学医学院罗素·H·摩根放射学与放射科学系，地址为马里兰州巴尔的摩市，邮编 21287（电子邮箱：jchen245@jhmi.edu）。

当年，TransUNet 的两个改进版本 TransUNet+[23] 和 Ds-TransUNet [24], 被提出，并在 CT分割任务中取得了更好的结果。对于三维分割而言，自注意力的计算成本会变得非常高昂，因此研究人员尝试限制 Transformer模块的使用，即仅在编码器和解码器网络之间的瓶颈层使用自注意力 [25, 26] ，或采用一种可变形机制，使注意力能够作用于少量关键位置 [27]。SegTran[28] 提出通过使用挤压与激励模块进行成对特征上下文化，利用更大上下文与定位精度之间的权衡。最近，越来越多的最先进性能由采用预训练Transformer骨干网络的网络刷新。预训练技术已成为Transformer领域新的研究方向，因为自注意力模块通常需要大规模预训练数据，才能学习到更强大的骨干网络 [29]。例如，自监督 Swin UNETR（Tang等人，2021 年）收集了大规模 CT图像（5,000 名受试者），用于预训练 Swin Transformer编码器，并在 BTCV [30] 和医学分割十项全能（MSD）[31]上取得了显著改进和最先进的性能。自监督掩码自编码器（MAE）[32] 研究了面向 Transformer的基于MAE的自预训练范式，通过从上下文中收集信息，强制网络预测被掩码的目标。除了开发先进架构以更好地学习数据之外，研究人员还尝试通过提供更具体匹配网络任务的额外数据来提升性能。

在表示学习中，多模态学习的发展使众多应用受益 [33, 34]。利用来自多种模态的融合特征，大幅提升了跨媒体分析任务的性能，例如视频分类 [35], 事件检测 [36, 37], 和情感分析 [38, 39]。这些研究共同展现出一个特点：如果在特征学习时提供多种模态（例如音频和视频），就能为单一模态（例如音频）学习到更优的特征。在 [40], Ngiam 等人提出了用于共享表示学习的跨模态（音频 + 视频）特征学习方案，并证明其视觉语音分类性能优于仅使用音频或仅使用视频数据训练的分类器。Wang 等人提出了一种基于深度神经网络的模型，将典型相关自编码器与基于自编码器的项相结合，以融合多视图，实现无监督的多视图特征学习 [41]。沿着这一趋势，基于深度学习的多模态方法也因在多项医学图像分析任务（包括分类 [42, 43], 诊断 [44, 45], 图像检索 [46], 和分割 [47-49]）中表现出色，而受到医学图像分析领域的关注。Carneiro 等人 [50] 提出利用同一区域未配准视图中的共享图像特征来提升分类性能。在 [44], Xu 等人提出利用多模态信息，联合学习图像与其他非图像模态之间的非线性相关性，以诊断宫颈发育不良；这一方法显著优于仅使用任一单一信息源的方法。在 [45], Suk 等人提出使用层次化深度卷积神经网络，从磁共振成像和正电子发射断层成像中学习联合特征表示，用于阿尔茨海默病诊断。

尽管视觉 Transformer 模型具有出色的表征能力，现有基于视觉 Transformer 的分割模型在接收多模态输入数据时，仍会产生不一致且不准确的密集预测。我们认为，这些模型的自注意力机制在提取多模态数据中存在的互补信息方面能力有限。为此，我们提出一种双分支交叉注意力 Swin Transformer（SwinCross），在不同尺度上融合来自两种不同模态的图像块，从而生成两种模态更具互补性的特征表示。此外，为降低计算量，我们基于交叉注意力和 Swin Transformer 的移位窗口自注意力机制，开发了一个跨模态注意力（CMA）模块 [10]。为验证所提方法的有效性，我们在一个公开数据集上进行了实验，并将该方法与 UNETR、Swin UNETR 和 nnU-Net 等最先进的方法进行了比较。实验结果表明，对于头颈部肿瘤分割任务，所提方法能够捕捉正电子发射断层成像与计算机断层扫描之间的模态间相关性。

II. 相关工作

A. 头颈部肿瘤分割的当前最先进方法

在 HECTOR 2021 挑战赛中，表现排名前五的团队在主要的头颈部肿瘤分割任务中均采用了 U-Net 或其变体 [51]。在 [52], Xie 等人采用基于图像块的 3D nnU-Net, 并结合压缩与激励归一化及一种新颖的训练方案，其中学习率通过 polyLR [53]动态调整。该方法在验证数据集上的五折平均 Dice 分数为 0.764，在肿瘤分割任务排行榜上名列第一。他们以五折交叉验证的方式训练了五个模型，并进行随机数据增强，包括旋转、缩放、镜像、高斯噪声和伽马校正。最终测试结果通过对五个测试预测结果进行概率平均集成得出。在 [54], An 等人提出了一种由粗到细框架，采用由三个 U-Net 组成的级联。第一个 U-Net 用于粗略分割肿瘤，然后选定一个边界框。接着，第二个 U-Net 在边界框内较小的区域进行更精细的分割，这通常能带来更准确的分割结果 [55]。最后一个 U-Net 将正电子发射断层成像、计算机断层扫描和先前的分割结果拼接后作为输入，以细化预测结果。三个 U-Net 使用不同的目标进行训练——第一个用于优化召回率，其余两个用于优化 Dice 系数。最终结果通过对三种不同的预测进行多数投票得出：由五个 nnU-Net 组成的集成模型、由三个采用压缩与激励（SE）归一化的 U-Net 组成的集成模型，以及所提出模型的预测结果。在 [56], Lu 等人提出了一种由十四个 3D U-Net 组成的大型集成学习模型，其中包括 [57], HECTOR 2020 挑战赛冠军采用的八个模型、五个通过留一中心法训练的模型，以及一个结合先验与

后验注意力的模型。最终的集成预测通过对全部十四个预测结果取平均，并将所得掩膜阈值化为 0.5 得出。在 [58], Yousefirizi 等人采用了经过留一中心法训练的 3D nnU-Net，并使用 SE 归一化，结合“统一”焦点损失与 Mumford-Shah损失，利用分布、区域和基于边界的损失函数的优势。最后，Ren 等人 [59] 提出了一种采用多种 PET归一化技术的 3D nnU-Net，即 PET-clip 和 PET-sin。前者对 [0,5] 中的标准摄取值（SUV）范围进行截断；后者则通过对 SUV 进行正弦变换，将空间上单调递增的 SUV 转换为洋葱环，因此在排行榜上排名第五。尽管基于CNN的方法具有出色的表征能力，但由于卷积核的感受野有限，它们缺乏对长程依赖关系进行建模的能力。这种感受野大小的固有限制，阻碍了全局语义信息的学习，而全局语义信息对于分割等密集预测任务至关重要。

B. Transformer与多模态学习 Transformer已广泛应用于

自然语言处理 [60—62] 和计算机视觉[63—68] 领域，这主要得益于其对长距离依赖关系的出色建模能力。除了在各种语言和视觉任务中取得令人瞩目的性能外，Transformer模型还通过将不同的模态输入作为词元进行自注意力[69—78]计算，为多模态推理提供了有效机制。例如，Prakash等人[74] 提出使用Transformer，通过注意力机制融合图像和激光雷达表示。超越语言和视觉领域，我们提出利用跨模态注意力 Swin Transformer，在多个分辨率下融合三维PET和CT图像，以实现头颈部肿瘤的分割。我们基于Swin Transformer的移位窗口块构建了SwinCross架构，该架构仅在局部区域内计算自注意力；相比之下，传统视觉Transformer的计算成本更高。尽管Swin Transformer无法显式计算超出其视野范围的对应关系，这在一定程度上类似于卷积网络的工作方式，但移位窗口机制仍能产生比大多数卷积网络大得多的卷积核 [79]。

III. SWINCROSS

A. SwinCross的整体架构

在这项工作中，我们提出了一种用于三维多模态分割的架构，包含两个主要组件：（1）用于整合多种模态（正电子发射断层成像和计算机断层扫描）信息的跨模态 Swin Transformer；（2）用于从不同模态中学习互补信息的跨模态移位窗口注意力块。我们的核心思路是，利用跨模态注意力机制，在头颈部肿瘤分割任务中，借助正电子发射断层成像和计算机断层扫描模态之间的互补性，将全局上下文纳入其中。图 1 展示了 SwinCross 的架构。输入到 SwinCross 模型的图像是一个多通道三维体积 $F^{in} \in R^{H\times W\times D\times M}$，其维度为 $H \times W \times D \times M$。首先，按通道拆分输入图像，得到一组单通道三维图像F_{mod_1}，…，$F_{mod_k} \in R^{H\times W\times D}$。然后，我们将每个单通道图像拆分为互不重叠的小图像块，图像块大小为 $\frac{H}{H'} \times \frac{W}{W'} \times \frac{D}{D'}$，对应的图像块分辨率为 $H' \times W' \times D'$。每个三维图像块都被投影到维度为 C 的嵌入空间中，形成词元序列 $S_{mod_k} \in R^{N\times C}$ $N = H' \times W' \times D'$，

其中 是序列中的词元数量，每个词元由一个维度为 C 的特征向量表示。S^{mod_K} 序列是编码器网络的输入。

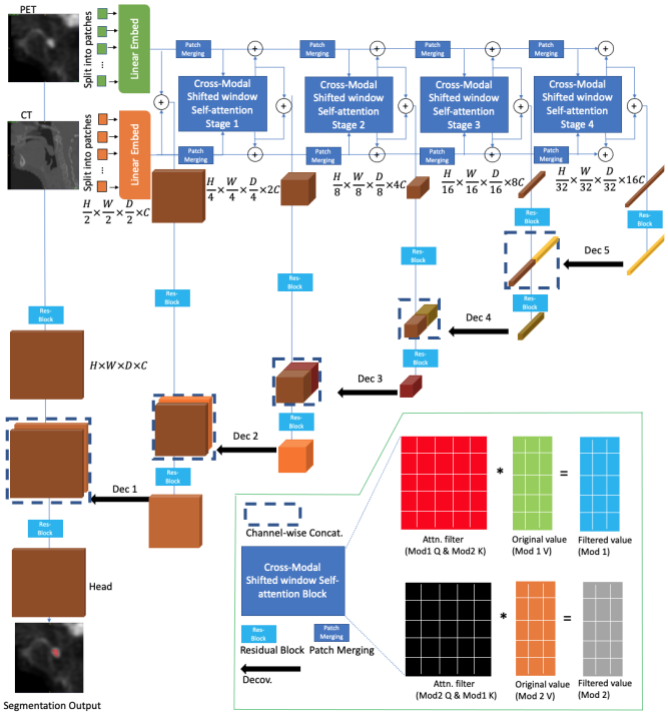

图1. SwinCross的架构。我们将三维PET和计算机断层扫描体数据作为跨模态注意力 Swin Transformer（SwinCross）的输入，该模型采用多个跨模态注意力（CMA）模块来融合两种模态的中间特征图。为了有效结合两种模态在不同尺度下的图像块词元，我们提出了一种基于CMA模块的融合方法，在两个特征提取分支中以多个分辨率（$\frac{1}{4}$、$\frac{1}{8}$、$\frac{1}{16}$、$and\frac{1}{32}$ 输入分辨率）交换信息，从而从两种模态中得到5个特征向量（$\frac{1}{2}$、$\frac{1}{4}$、$\frac{1}{8}$、$\frac{1}{16}$、$and\frac{1}{32}$ 输入分辨率），并通过逐元素求和将它们结合起来。这5个特征向量构成了CT和PET图像在5种不同分辨率下的融合表示。随后，卷积网络解码器对这些特征向量进行处理，并预测最终的分割图。我们将上一分辨率解码后的特征向量与当前分辨率的特征向量按通道拼接，并将所得特征向量作为反卷积模块的输入，以生成下一分辨率下的特征向量。

B. 网络编码器

编码器使用线性投影，为每个输入序列 S_{mod_k}计算一组查询、键

$$Q_{mod_k} = S_{mod_k}M^q, K_{mod_k} = S_{mod_k}M^k, V_{mod_k} = S_{mod_k}M^v \quad (1)$$

其中 $M^q \in R^{D_f \times D_q}$、$M^k \in R^{D_f \times D_k}$ 和 $M^v \in R^{D_f \times D_v}$ 是权重矩阵。在双模态交叉注意力中，它使用每种模态的 Q 和 K 之间的缩放点积来计算注意力权重，然后聚合每种模态中每个查询对应的值，

$$A_{mod_1} = softmax\left(\frac{Q_{mod_1} K_{mod_2}^T}{\sqrt{D_k}}\right) V_{mod_1}, \quad (2)$$

$$A_{mod_2} = softmax\left(\frac{Q_{mod_2} K_{mod_1}^T}{\sqrt{D_k}}\right) V_{mod_2}, \quad (3)$$

其中，Q_{mod_1}, K_{mod_1}, V_{mod_1}, Q_{mod_2}, K_{mod_2}, V_{mod_2} 分别表示来自模态 1 和模态 2 的查询、键和值；D_k 表示键和查询的大小。

由于这些是三维令牌，且注意力计算的成本随令牌数量呈平方增长，因此我们采用移位窗口机制进行交叉注意力计算。具体而言，在 Transformer 编码器的给定层 l 中，我们使用大小为 $M \times M \times M$ 的窗口，将分块体积均匀划分为 ^{H},× ^{W},×$_{M}$ $_{M}^{D}$,个区域。在编码器后续的 $_{M}$层 l 和 $\overline{l}+1$ 中，输出计算如下：

$$\hat{A}_{mod_k}^{l} = \text{W-MSA}\left(\text{LN}\left(A_{mod_k}^{l-1}\right)\right) + A_{mod_k}^{l-1} \quad (4)$$

$$A_{mod_k}^{l} = \text{MLP}\left(\text{LN}\left(\hat{A}_{mod_k}^{l}\right)\right) + \hat{A}_{mod_k}^{l} \quad (5)$$

$$\hat{A}_{mod_k}^{l+1} = \text{SW-MSA}\left(\text{LN}\left(A_{mod_k}^{l}\right)\right) + A_{mod_k}^{l} \quad (6)$$

$$A_{mod_k}^{l+1} = \text{MLP}\left(\text{LN}\left(\hat{A}_{mod_k}^{l+1}\right)\right) + \hat{A}_{mod_k}^{l+1} \quad (7)$$

我们实现了循环移位 [10] 的三维版本，以高效计算移位窗口机制。SwinCross采用标准的四阶段结构 [10]，但在每个阶段都使用跨模态注意力机制，融合两种模态的中间特征图。特征提取器在两种模态的多个分辨率（$\frac{H}{2}\times\frac{W}{2}\times\frac{D}{2}\times C, \frac{H}{4}\times\frac{W}{4}\times\frac{D}{4}\times 2C, \frac{H}{8}\times\frac{W}{8}\times\frac{D}{8}\times 4C,$ $\frac{H}{16}\times\frac{W}{16}\times\frac{D}{16}\times 8C$）上进行融合，从而使每种模态都得到四个经过滤波的特征图（$\frac{H}{4}\times\frac{W}{4}\times\frac{D}{4}\times 2C, \frac{H}{8}\times\frac{W}{8}\times\frac{D}{8}\times 4C,$ $\frac{H}{16}\times\frac{W}{16}\times\frac{D}{16}\times 8C, \frac{H}{32}\times\frac{W}{32}\times\frac{D}{32}\times 16C$）。两种模态经过滤波的特征图逐元素求和后送入解码器，如图 1 中红色加号所示。在每个阶段，这些特征图都会与下采样（通过分块合并）后的输入特征图逐元素求和，再反馈到各自的模态分支中，如图 1 中绿色加号所示。

SwinCross 编码器的图像块大小为 $2 \times 2 \times 2$，特征维度为 $2 \times 2 \times 2 \times 2 = 16$，这考虑了具有 2 个通道的多模态 PET/CT 图像。嵌入空间的大小 C 在我们的编码器中设为 48。此外，SwinCross 编码器包含 4 个阶段，每个阶段由 [2,4,2,2] 个跨模态移位窗口变换器块组成。因此，编码器中的总层数为 L = 10。在阶段 1 之前，使用步长为 2（即图像块大小为 2）、输出通道数为 C 的三维卷积层，将每个单通道图像拆分为互不重叠的小图像块，从而得到 $\frac{H}{}\times\frac{W}{}\times\frac{D}{}\times C$ 3D 个词元。为了遵循 [10], 中提出的 $_{2}$ $_{22}$ 层次结构，在每个阶段开始时，都会对各模态分支应用图像块合并层，使特征表示的分辨率降低 2 倍。为了将输入图像中的精细细节保留到输出分割中，我们将原始多通道三维输入体积 $F^{in} \in R^{H \times W \times D \times M}$ 、其嵌入版本以及 4 个阶段输出的特征图一并送入解码器，最终得到总计 6 个特征图，其维度分别为 $H \times W \times D \times M$、$\frac{H}{}\times\frac{W}{}\times\frac{D}{}\times C$、$\frac{H}{}\times\frac{W}{}\times\frac{D}{}\times$ $_{2}$ $_{2}$ $_{2}$ $_{4}$ $_{44}$ $2C$ $\frac{H}{8}\times\frac{W}{8}\times\frac{D}{8}\times 4C$ $\frac{H}{16}\times\frac{W}{16}\times\frac{D}{16}\times 8C$ $\frac{H}{32}\times\frac{W}{32}\times\frac{D}{32}\times$、、 $16C$ 、和。

C. 网络解码器

为便于跨模态特征融合并降低计算成本，我们采用卷积网络解码器，而非变换器解码器。SwinCross采用U形网络设计，在每个分辨率下，编码器提取的特征表示通过跳跃连接传递至解码器。在编码器的每个阶段i ($i \in [0,1,2,3,4,5]$)，输出特征表示会被重塑为大小 $\frac{H}{2^i}\times\frac{W}{2^i}\times\frac{D}{2^i}$，并输入一个残差块。该残差块由两个3x3x3卷积层组成，并通过实例归一化层进行归一化。随后，使用反卷积层将特征图的分辨率提高2倍，并将其输出与前一阶段的输出拼接。然后，将拼接后的特征输入另一个如前所述的残差块。最后，通过一个1x1x1卷积层和Sigmoid激活函数计算分割结果。

IV. 结果与讨论

A. HECTOR 2021年数据集上的消融研究

在表 1 中，我们对 CMA 模块块进行消融实验，该模块块仅涉及 Swin Transformer 的注意力机制，其他设置均与 Swin UNETR 保持一致（例如，嵌入维度、特征大小、各阶段的块数、窗口大小和头数）。我们从按通道拼接的输入开始，该输入由来自两种模态的两幅体积图像组成。这种多模态输入已使 Swin UNETR 获得 0.754±0.032 的五折平均 Dice 评分。如果将两种模态的图像分别输入两个分支（如图 1 所示），并使用 CMA 模块块在每个阶段融合各模态学习到的特征，则性能提升至 0.769±0.026。每个 CMA 模块块的输出形状与输入相同，每个经过滤的特征都会加回到相应模态的分支中。在每个阶段，

CMA 模块块输出的经过滤的特征之和会被送入解码器。

表 1. HECKTOR 2021年数据集上的 CMA 模块消融研究

模型	块组成	嵌入维度 on	特征 Size	数量 of 块	窗口 Size	数量 头部	五折平均 ge Dice 分数
Swin UNETR	W-MSA+ SW-MSA	768	48	[2,2,2,2]	[7,7,7]	[3,6,12,24]	0.754 ±0.03 2
	CMW-MSA + CMSW-MSA	768	48	[2,2,2,2]	[7,7,7]	[3,6,12,24]	**0.769 ±0.02 6**

B. 与医学图像分割领域最先进的方法的比较 我们将 SwinCross 的性能与

当前医学图像分割领域的最先进方法，如 Swin UNETR、UNETR 和 nnU-Net，均采用五折交叉验证划分。表 2 展示了所有五折的评估结果（双模态）。在所有对比方法中，所提出的 SwinCross 模型取得了最高的五折平均Dice评分，为 0.769。值得注意的是，SwinCross 在全部五折中均优于 Swin UNETR，这表明它能够通过跨模态注意力模块，在多个分辨率下学习多模态特征表示。这些结果与我们先前的发现一致 [80]，，即我们曾指出，在两个公开数据集上进行头颈部肿瘤分割时，nnU-Net 的表现优于 Swin UNETR。借助 CMA模块和双分支融合机制，SwinCross 的分割表现略优于 nnU-Net，五折平均Dice评分也体现了这一点。不过，nnU-Net 仍展现出有竞争力的表现，这再次表明，对于小目标分割，建模长程依赖带来的提升可能有限，因为较小的有效感受野或许就足以捕获小目标（如头颈部肿瘤）中的全部前景和背景信息 [80]。

表 2. 使用正电子发射断层成像和 CT图像的所有方法的平均 Dice 系数值对应的五折交叉验证基准。

Dice系数	SwinCross (正电子发射断层成像+计算机断层扫描) 本文提出的方法	Swin UNETR (正电子发射断层成像+计算机断层扫描)	nnU-Net (正电子发射断层成像+计算机断层扫描)	UNETR (正电子发射断层成像+计算机断层扫描)
第0折	**0.717**	0.715	0.714	0.702
第1折	**0.788**	0.781	0.781	0.716
第2折	0.800	0.752	**0.803**	0.727
第3折	**0.779**	0.772	0.777	0.762
第4折	**0.761**	0.748	**0.761**	0.708
平均值	**0.769**	0.754	0.767	0.723

图2展示了所有方法的样本分割结果。对于较大的肿瘤，图2a显示了基于 Transformer 的模型通过建模长程依赖所带来的优势。SwinCross 是唯一能够捕捉到黄色十字准线所标记的肿瘤尖端的网络，这证明了 CMA模块块的优势；这些模块能够在编码器中的多个分辨率下实现两种模态之间的特征交换。对于较小的肿瘤，图2c显示了 SwinCross 能够通过融合来自 CT图像的互补边缘特征来捕获肿瘤的精细边缘，表现优于采用按通道拼接的输入的其他方法。结果证实了 [34] 中的发现——在网络内部进行融合优于在网络外部进行融合。

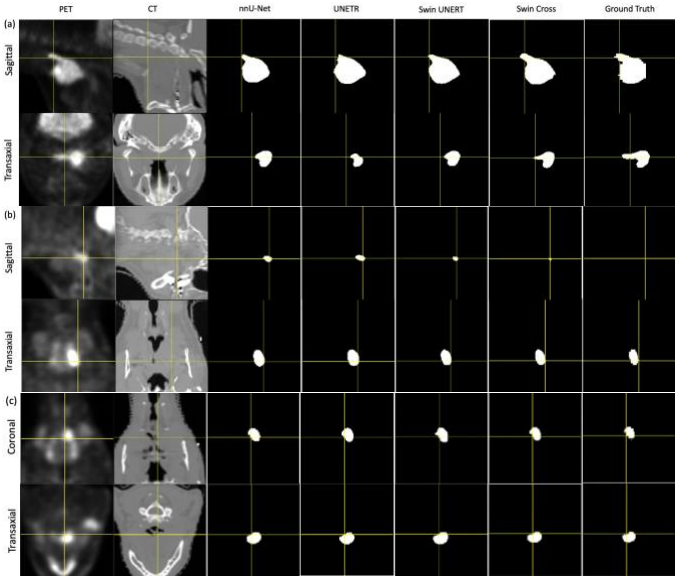

图2. 从左至右依次为输入PET图像、CT图像，以及nnU-Net、UNETR、Swin UNETR、SwinCross（本文提出）推理得到的掩膜和真实标注。

C. 与单模态分割的比较

作为与使用双模态输入的比较，我们使用相同的五折交叉验证划分，计算了参考方法在单模态输入下的性能。由于头颈部肿瘤主要存在于PET中，因此我们仅使用PET图像开展了这些单模态实验。表3展示了所有五折中的评估结果（仅PET）。在单模态输入的参考方法中，Swin UNETR取得了最高的五折平均 Dice 系数，为0.732。图3展示了所有方法在使用单模态和双模态时，五个划分所得平均 Dice 系数值的箱线图。总体而言，该图表明，对于头颈部肿瘤分割任务，采用双模态（PET和CT）输入的网络明显优于采用单模态（PET）输入的相同网络。

表3. 仅使用PET图像时，各参考方法在平均 Dice 系数值方面的五折交叉验证基准。

Dice系数	UNETR （正电子发射断层成像）	Swin UNETR （正电子发射断层成像）	nnU-Net （正电子发射断层成像）
第0折	0.668	0.689	0.683
第1折	0.677	0.711	0.701
第2折	0.715	0.768	0.762
第3折	0.722	0.762	0.745
第4折	0.686	0.729	0.728
平均值	0.694	0.732	0.724

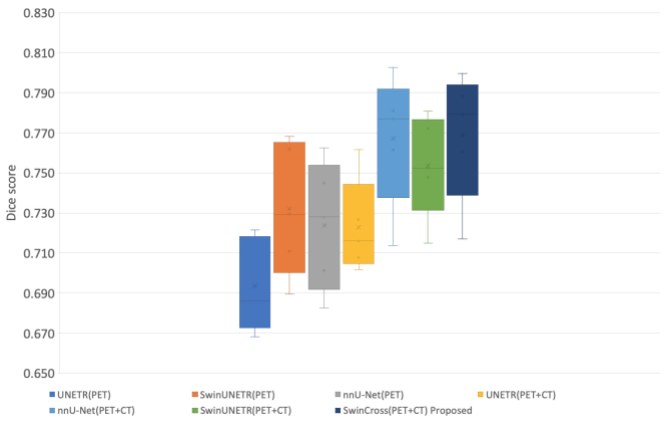

图3. 不同方法在五个划分中的平均 Dice 系数值箱线图。

V. 结论

我们提出了一种跨模态 Swin Transformer，用于在PET和CT图像中自动勾画头颈部肿瘤。所提出的模型包含一个跨模态注意力模块，该模块在多个分辨率下利用两种模态之间的特征交换。基于卷积网络的解码器通过不同分辨率下的跳跃连接与编码器相连。我们使用HECTOR 2021数据集，通过与最先进的方法进行比较，验证了所提出模型的有效性。实验表明，所提出的方法能够更好地捕捉PET和CT之间的模态间相关性，在头颈部肿瘤分割任务中的表现优于其他方法。所提出的方法通常也适用于使用其他成像模态（如SPECT/CT或磁共振成像）的其他语义分割任务。

参考文献

1. Fabian Isensee、P.F.J.、Simon A. A. Kohl、Jens Petersen 和 Klaus H. Maier-Hei，《nnU-Net：一种用于基于深度学习的生物医学图像分割的自配置方法》。《自然·方法》2020年。2. Ali Hatamizadeh、Y.T.、Vishwesh Nath、Dong Yang、Andriy Myronenko、Bennett Landman、Holger R. Roth、Daguang Xu。《UNETR：用于三维医学图像分割的 Transformer》。载于 IEEE/CVF冬季计算机视觉应用会议（WACV）。2022年。3. Ali Hatamizadeh、V.N.、Yucheng Tang、Dong Yang、Holger R. Roth 和 Daguang Xu，《Swin UNETR：用于脑肿瘤磁共振成像图像语义分割的 Swin Transformer》。arXiv，2022年。4. Parkin, D.M. 等，《全球癌症统计，2002年》。《CA：临床癌症杂志》，2005年，55(2)：第74–108页。5. Andrearczyk, V. 等。HECKTOR挑战赛概览（医学图像计算与计算机辅助干预会议2020）：PET/CT中的头颈部肿瘤自动分割。2021年。Cham：Springer International Publishing。

6. J. Li 等，Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. arXiv 预印本 arXiv:2206.01136，2022年。7. L. Yuan、Q.H.、Z. Jiang、J. Feng 和 S. Yan，VOLO：用于视觉识别的视觉展望器。IEEE模式分析与机器智能汇刊，2022年。8. Yang、Yilin 和 Jianming Zhang、He Zhang、Zijun Wei、Zhe Lin、Alan Yuille。具有增强自注意力的轻量级视觉Transformer。载于IEEE/CVF计算机视觉与模式识别会议。2022年。9. Wang、Enze、Xiang Li、Deng-Ping Fan、Kaitao Song、Ding Liang、Tong Lu、Ping Luo 和 Ling Shao，Pvtv2：采用金字塔视觉Transformer的改进基线。《计算视觉媒体》，2022年。10. Z. Liu 等。Swin Transformer：使用移位窗口的层次化视觉Transformer。载于 IEEE/CVF国际计算机视觉会议论文集。2021年。11. Bowen Cheng、A.S.、Alexander Kirillov。逐像素分类并非语义分割所需的一切。载于神经信息处理系统进展。2021年。12. nze Xie、W.W.、Zhiding Yu、Anima Anandkumar、Jose M. Alvarez、Ping Luo。SegFormer：使用 Transformer 进行语义分割的简洁高效设计。载于神经信息处理系统进展。2021年。13. Luo、Z.L.、W.W.、E.X.、Z.Y.、A.A.、J.M.A.、T.L.、P. 全景分割SegFormer：深入探索使用 Transformer 的全景分割。载于 IEEE/CVF计算机视觉与模式识别会议。2022年。14. Wan、Z.、Jingbo、Dongdong Chen 和 Jing Liao。使用 Transformer 实现高保真多样化图像补全。载于 IEEE/CVF计算机视觉与模式识别会议。2022年。15. Wang、H.、Yukun、Hartwig Adam、Alan Yuille 和 Liang-Chieh Chen。MaX-DeepLab：使用掩码Transformer进行端到端全景分割。载于 IEEE/CVF计算机视觉与模式识别会议。2022年。16. B. Hou 等。Ratchet：用于胸部X线诊断和报告的医学 Transformer。载于《医学图像计算与计算机辅助干预–MICCAI 2021：第24届国际会议，法国斯特拉斯堡，2021年9月27–10月1日，会议论文集，第七部分》24。2021年。Springer。17. C. Matsoukas 等，是否到了用 Transformer 取代医学图像卷积神经网络的时候？arXiv 预印本 arXiv:2108.09038，2021年。18. S. Park 等，用于新冠肺炎胸部X线诊断的联邦式分割任务无关视觉Transformer。神经信息处理系统进展，2021年。34：第24617–24630页。

19. Chen、Yongyi、Yu、Qihang、Luo、32. Xiangde、Adeli, Ehsan、Wang, Yan、Lu, Le、Alan Yuille 和 Zhou, Yuyin，TransUNet：Transformer 为医学图像分割构建强大的编码器。Arxiv.org，2021年。20. A. Hatamizadeh 等。Swin UNETR：用于脑肿瘤磁共振成像图像语义分割的 Swin Transformer。载于《脑病灶：胶质瘤、多发性硬化症、中风和创伤性脑损伤：第七届国际研讨会，BrainLes 2021，与 MICCAI 2021 联合举办，线上活动，2021年9月27日，修订精选论文，第一部分》。2022年。Springer。21. J. Chen 等，TransMorph：用于无监督医学图像配准的 Transformer。《医学图像分析》，2022年。82：第102615页。22. J. Chen 等，ViT-V-Net：用于无监督体积医学图像配准的视觉 Transformer。arXiv 预印本 arXiv:2104.06468，2021年。23. Yuhang Liu、H. Wang、Zugang Chen、Kehan Huangliang、Haixian Zhang，TransUNet+：重新设计跳跃连接以增强医学图像分割中的特征。《Knowledge-Based Systems》，2022年。256。24. Lin、A.、Bingzhi、Jiayu Xu、Zhang, Zheng 和 Lu, Guangming，DS-TransUNet：用于医学图像分割的双路 Swin Transformer U-Net。Arxiv.org，2021年。25. Chang 等，TransClaw U-Net：结合 Transformer 的 Claw U-Net，用于医学图像分割。Arxiv，2021年。abs/2107.05188。26. Deng、K.、Yanda、Dongxu Gao、Joshua Bridge、Yaochun Shen、Gregory Lip、Yitian Zhao 和 Yalin Zheng，TransBridge：用于超声心动图左心室分割的轻量级 Transformer，载于《简化医学超声：第二届国际研讨会，ASMUS 2021，与 MICCAI 2021 联合举办，法国斯特拉斯堡，2021年9月27日，会议论文集》。2021年。第63–72页。27. Xie、Y.、Jianpeng、Chunhua Shen 和 Yong Xia。CoTr：高效连接 CNN 与 Transformer，用于三维医学图像分割。载于医学图像计算与计算机辅助干预国际会议。2021年。28. S. Li 等，使用挤压与扩展 Transformer 进行医学图像分割。Arxiv，2021年。abs/2105.09511。29. A. Dosovitskiy 等，一张图像胜过16×16个词：大规模图像识别 Transformer。arXiv 2020。arXiv 预印本 arXiv:2010.11929，2010年。30. B. Landman 等。MICCAI 颅腔外多图谱标注–研讨会与挑战赛。载于《MICCAI 颅腔外多图谱标注—研讨会与挑战赛论文集》。2015年。31. M. Antonelli 等，《医学分割十项全能》。《Nature Communications》，2022年。13(1)：第4128页。

33. Zhou, L. 等，自监督预训练与掩码自编码器在医学图像分析中的应用。arXiv 预印本 arXiv:2203.05573，2022年。

34. Guo, W.Z.、J.W. Wang 和 S.P. Wang，深度多模态表示学习：综述。IEEE Access，2019年。7：第 63373-63394 页。

35. Guo, Z. 等，基于深度学习的多模态医学成像图像分割。IEEE Trans Radiat Plasma Med Sci，2019年。3(2)：第 162-169 页。

36. Liu, Y.、X. Feng 和 Z. Zhou，使用堆叠收缩自编码器进行多模态视频分类。Signal Processing，2016年。120：第 761-766 页。

37. Wu, S. 等。利用弱监督概念的多模态融合进行零样本事件检测。载于 IEEE计算机视觉与模式识别会议论文集。2014。

38. Habibian, A.、T. Mensink 和 C.G. Snoek, Video2vec 嵌入在示例稀少时识别事件。IEEE模式分析与机器智能汇刊，2016。39(10)：第 2089-2103 页。

39. Poria, S. 等，融合音频、视觉和文本线索，从多模态内容中进行情感分析。神经计算，2016。174：第 50-59 页。

40. Zadeh, A. 等，用于多模态情感分析的张量融合网络。arXiv 预印本 arXiv:1707.07250，2017。

41. Ngiam, J. 等。多模态深度学习。载于国际机器学习会议。2011。

42. Wang, W. 等，《深度多视图表示学习：目标与优化》。ArXiv，2016。abs/1602.01024。

43. Zhu, Z.Y. 等。多视图感知器：用于学习人脸身份和视图表示的深度模型。载于第28届神经信息处理系统会议（NIPS）。2014年。加拿大蒙特利尔。

44. Carneiro, G.、J. Nascimento 和 A.P. Bradley。利用预训练深度学习模型进行未配准多视图乳腺摄影图像分析。载于第18届医学图像计算与计算机辅助干预国际会议（医学图像计算与计算机辅助干预会议）。2015年。德国慕尼黑。

45. Xu, T. 等。用于宫颈发育不良诊断的多模态深度学习。载于医学图像计算与计算机辅助干预会议。2016年。

46. Suk, H.I. 等，利用深度学习进行层次化特征表示和多模态融合，以实现阿尔茨海默病/轻度认知障碍诊断。Neuroimage，2014年。101：第569–582页。

47. Kang, Y.、S. Kim 和 S. Choi。利用多重表示进行深度学习哈希。载于第12届IEEE国际数据挖掘会议（ICDM）。2012年。比利时布鲁塞尔。

48. Zhu, X. 等，通过将多模态图像与模态加权UNet相结合进行医学病灶分割。Medical Physics，2022年。49(6)：第3692—3704页。

Guo, Z. 等，《基于多模态卷积神经网络的医学图像分割：图像融合方案研究》。载于第15届IEEE国际生物医学成像研讨会（ISBI）。2018年。华盛顿特区。

49. Guo, Z. 等，使用深度稠密多模态网络进行头颈癌放疗的大体肿瘤体积分割。Phys Med Biol，2019年。64(20)：第205015页。50. Carneiro, G.、J. Nascimento 和 A.P. Bradley。使用预训练深度学习模型分析未配准多视图乳腺摄影图像。载于医学图像计算与计算机辅助干预国际会议。2015年。Springer。51. Andrearczyk, V. 等。《MICCAI 2021 HECKTOR挑战赛综述：PET/CT图像中的头颈部肿瘤自动分割与结局预测》。2022年。瑞士湛：Springer International Publishing。52. Xie, J. 和 Y. Peng。《基于三维U-Net的头颈部肿瘤分割》。2022年。瑞士湛：Springer International Publishing。53. Chen, L.-C. 等，《Deeplab：利用深度卷积网络、空洞卷积和全连接条件随机场进行语义图像分割》。IEEE模式分析与机器智能汇刊，2017年。40(4)：第834–848页。54. An, C.、H. Chen 和 L. Wang。《用于CT和PET图像中头颈部肿瘤分割的由粗到细框架》。2022年。瑞士湛：Springer International Publishing。55. Zhou, Y. 等。《用于腹部CT扫描中胰腺分割的不动点模型》。2017年。瑞士湛：Springer International Publishing。56. Lu, J. 等。《用于泛化头颈部肿瘤分割的先验与后验注意力》。2022年。瑞士湛：Springer International Publishing。57. Iantsen, A.、D. Visvikis 和 M. Hatt。《用于PET与CT融合图像中头颈部原发肿瘤自动勾画的挤压与激励归一化》。2021年。瑞士湛：Springer International Publishing。58. Yousefirizi, F. 等。《使用三维U-Net和Cox比例风险神经网络对PET/CT体积数据中的头颈癌进行分割和风险评分预测》。2022年。瑞士湛：Springer International Publishing。59. Ren, J. 等。《PET归一化：提升三维PET/CT中头颈部肿瘤的深度学习自动分割效果》。2022年。瑞士湛：Springer International Publishing。60. Vaswani, A. 等。《Attention Is All You Need》。载于第31届神经信息处理系统年会（NIPS）。2017年。加利福尼亚州长滩。61. Devlin, J. 等，《Bert：用于语言理解的深度双向变换器预训练》。arXiv预印本 arXiv:1810.04805，2018年。62. Brown, T. 等，《语言模型是少样本学习器》。神经信息处理系统进展，2020年。33：第1877–1901页。63. Wang, X. 等。《非局部神经网络》。载于IEEE计算机视觉与模式识别会议论文集。2018年。

65. Parmar, N. 等。图像 Transformer。载于国际机器学习会议。2018年。PMLR。

66. Dosovitskiy, A. 等，图像胜过千言万语：大规模图像识别 Transformer。arXiv 预印本 arXiv:2010.11929，2020年。

67. Carion, N. 等。基于 Transformer 的端到端目标检测。载于欧洲计算机视觉会议。2020年。Springer。

68. Chen, M. 等，《像素生成式预训练》，载于《第37届国际机器学习会议论文集》，D. Hal, III 和 S. Aarti 编。2020年，PMLR：《机器学习研究论文集》，第1691--1703页。

69. Chen, C.-F.R.、Q. Fan 和 R. Panda。《Crossvit：用于图像分类的跨注意力多尺度视觉 Transformer》。载于《IEEE/CVF国际计算机视觉会议论文集》。2021年。

70. Tan, H. 和 M. Bansal，《Lxmert：从 Transformer 中学习跨模态编码器表示》。arXiv 预印本 arXiv:1908.07490，2019年。

71. Li, L.H. 等，《Visualbert：视觉与语言的简单高性能基线》。arXiv 预印本 arXiv:1908.03557，2019年。

72. Sun, C. 等。《Videobert：用于视频与语言表示学习的联合模型》。载于《IEEE/CVF国际计算机视觉会议论文集》。2019年。

73. Chen, Y.-C., 等. Uniter：通用图像文本表示学习. 载于欧洲计算机视觉会议. 2020年. Springer.

74. Li, X., 等. Oscar：面向视觉语言任务的对象语义对齐预训练. 载于欧洲计算机视觉会议. 2020年. Springer.

75. Prakash, A., K. Chitta, 和 A. Geiger. 用于端到端自动驾驶的多模态融合Transformer. 载于IEEE/CVF计算机视觉与模式识别会议论文集. 2021年.

76. Huang, Z., 等. 跳出框架的视觉感知：面向视觉语言表示学习的端到端预训练. 载于IEEE/CVF计算机视觉与模式识别会议论文集. 2021年.

77. Hu, R. 和 A. Singh. Unit：借助统一Transformer进行多模态多任务学习. 载于IEEE/CVF国际计算机视觉会议论文集. 2021年.

78. Akbari, H., 等，Vatt：利用原始视频、音频和文本进行多模态自监督学习的Transformer，神经信息处理系统进展，2021年。34：第24206-24221页.

79. Hendricks, L.A. 等，解耦数据、注意力和损失函数在多模态 Transformer 中的作用。《计算语言学协会会刊》，2021年。9：第570-585页。

80. Ding, X. 等，将卷积核扩展至31×31：重新审视卷积神经网络中的大卷积核设计。发表于《IEEE/CVF计算机视觉与模式识别会议论文集》。2022年。

Ye Li、J.C.、Se-in Jang、Kuang Gong、Quanzheng Li，《多模态网络架构研究》

《头颈部肿瘤分割》。Arxiv.org，2022年。