Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation

Hu Cao

摘要

在医学图像分析中，卷积神经网络（CNNs）在过去几年中取得了显著的成就。尤其是基于 U 形架构和跳跃连接的深度神经网络在各种医学图像任务中被广泛应用。然而，尽管 CNN 在性能上取得了卓越的成就，但由于卷积操作的局部性，它无法很好地学习全局和长距离的语义信息交互。因此，文章提出了 Swin-Unet，这是一种基于 Transformer 的纯粹 U 型结构，用于医学图像分割。文章采用分词的图像块输入到基于 Transformer 的 U 形编码器-解码器架构中，该架构具有跳跃连接以进行局部-全局语义特征学习。具体而言，文章使用具有偏移窗口的分层 Swin Transformer 作为编码器来提取上下文特征。并设计了基于 Swin Transformer 的对称解码器，其中包含用于执行上采样操作以恢复特征图的空间分辨率的块扩展层。在输入和输出的直接 4× 下采样和上采样的情况下，实验证明了基于纯 Transformer 的 U 型编码器-解码器网络在多器官和心脏分割任务中优于全卷积或 Transformer 与卷积组合的方法。

关键词：医学图像分割；Swin-Unet；Transformer

1 引言

已有的医学图像分割方法主要依赖于具有 U 型结构的完全卷积神经网络（FCNN）[1][2][3]。典型的 U 型网络，U-Net[1]，由对称的编码器-解码器和跳跃连接组成。在编码器中，使用一系列卷积层和连续的下采样层提取具有较大感受野的深层特征。然后，解码器对提取的深层特征进行上采样，以进行像素级语义预测，并通过跳跃连接将来自编码器的不同尺度的高分辨率特征融合，以减轻由于下采样而导致的空间信息丢失。通过这种优雅的结构设计，U-Net 在各种医学图像应用中取得了巨大成功。沿着这个技术路径，许多算法如 3D U-Net[4]、Res-UNet[5]、U-Net++[6]和 UNet3+[7]已经为各种医学图像模态的图像和体积分割开发出来。这些基于 FCNN 的方法在心脏分割、器官分割和病变分割等方面取得了出色的性能，证明了 CNN 在学习判别特征方面的强大能力。

然而，尽管基于 CNN 的方法在医学图像分割领域取得了出色的性能，它们仍然无法完全满足对分割准确性的严格要求。由于卷积操作的固有局部性，基于 CNN 的方法很难学习明确的全局和长程语义信息交互。一些研究尝试通过使用空洞卷积层[8][9]、自注意机制[10][11]和图像金字塔[12]来解决这个问题。然而，这些方法在建模长程依赖性方面仍然存在局限。受 Transformer 在自然语言处理（NLP）领域的巨大成功[13]启发，研究人员尝试将 Transformer 引入视觉领域[14]。在[15]中，提出了视觉 Transformer（ViT）来执行图像识别任务。采用 2D 图像块和位置嵌入作为输入，并在大型数据集上进行预训练，ViT 在性能上与基于 CNN 的方法相媲美。此外，[16]中提出了数据高效图像 Transformer（DeiT），表明 Transformer 可以在中等规模的数据集上训练，并通过与蒸馏方法相结合获得更强大的 Transformer。在[17]中，开发了分层 Swin Transformer。将 Swin Transformer 作为视觉骨干，[17]的作者在图像分类、目

标检测和语义分割方面取得了最先进的性能。ViT、DeiT 和 Swin Transformer 在图像识别任务中的成功表明，Transformer 有望应用于视觉领域。

在受 Swin Transformer[17]成功的启发下，本文中提出了 Swin-Unet，以利用 Transformer 在 2D 医学图像分割中的潜力。据所知，Swin-Unet 是第一个完全基于 Transformer 的 U 型结构，包括编码器、瓶颈、解码器和跳跃连接。编码器、瓶颈和解码器都是基于 Swin Transformer 块[17]构建的。将输入的医学图像分割成不重叠的图像块，每个块被视为一个令牌，并输入到基于 Transformer 的编码器中，以学习深层特征表示。通过带有图块扩展层的解码器进行上采样操作，通过跳跃连接与来自编码器的多尺度特征融合，从而恢复特征图的空间分辨率，并进一步进行分割预测。对多器官和心脏分割数据集的广泛实验证明，该方法具有出色的分割准确性和鲁棒的泛化能力。

2 相关工作

2.1 基于传统 CNN 的方法

早期的医学图像分割方法主要是基于轮廓和传统的基于机器学习的算法[18][19]。随着深度 CNN 的发展，U-Net 在[1]中被提出用于医学图像分割。由于 U 型结构的简单性和卓越性能，不断涌现出各种类似 Unet 的方法，如 Res-UNet[5]、Dense-UNet[20]、U-Net++[6]和 UNet3+[7]。它还被引入到 3D 医学图像分割领域，如 3D-Unet[4]和 V-Net[21]。目前，基于 CNN 的方法在医学图像分割领域取得了巨大的成功，这归功于其强大的表示能力。

2.2 视觉 Transformer

Transformer 首次在[13]中被提出，用于机器翻译任务。在自然语言处理（NLP）领域，基于 Transformer 的方法在各种任务中取得了最先进的性能[22]。在 Transformer 的成功推动下，研究人员在[15]中引入了一种先进的视觉 Transformer（ViT），在图像识别任务中取得了令人印象深刻的速度-精度平衡。与基于 CNN 的方法相比，ViT 的缺点是需要在自己的大型数据集上进行预训练。为了缓解 ViT 训练的困难，Deit[16]描述了几种训练策略，使 ViT 能够在 ImageNet 上训练良好。最近，一些基于 ViT 的优秀工作已经完成[23][24][17]。值得一提的是，一种高效且有效的层次视觉 Transformer，称为 Swin Transformer，在[17]] 中被提出作为视觉骨干。基于移位窗口机制，Swin Transformer 在包括图像分类、目标检测和语义分割在内的各种视觉任务上取得了最先进的性能。在这项工作中，本文尝试使用 Swin Transformer 块作为基本单元，构建具有跳跃连接的 U 型编码器-解码器结构，用于医学图像分割，从而为 Transformer 在医学图像领域的发展提供基准比较。

3 本文方法

3.1 本文方法概述

Swin-Unet 的基本单元是 Swin Transformer 块，其中编码器通过将医学图像分割成非重叠的 4×4 大小的补丁，然后通过 Swin Transformer 块和补丁合并层生成分层特征表示。解码器使用 Swin Transformer 块和补丁扩展层，通过跳跃连接将来自编码器的上下文特征与解码器的多尺度特征进行融合，以弥补由下采样导致的空间信息丢失。最后，通过补丁扩展层执行上采样，将特征图的分辨率恢复到输入分辨率，并通过线性投影层输出像素级分割预测。整体架构如图所示：

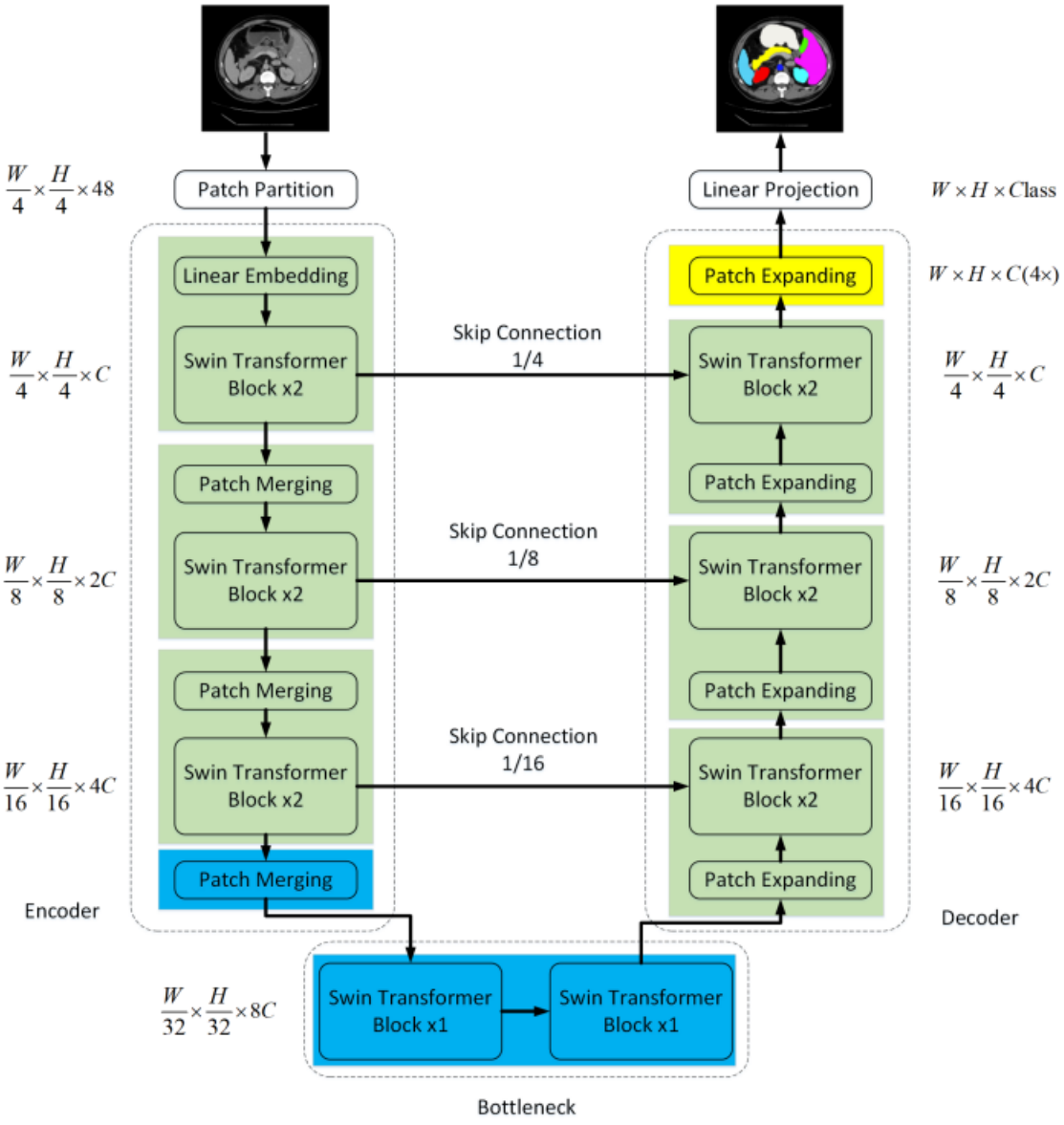

图 1: swin-unet 架构图

3.2 Swin Transformer 模块

与传统的多头自注意力（MSA）模块不同，Swin Transformer 块[17]是基于平移窗口构建的。在图 2 中，展示了两个连续的 Swin Transformer 块。每个 Swin Transformer 块由 LayerNorm（LN）层、多头自注意力模块、残差连接和具有 GELU 非线性的 2 层 MLP 组成。分别在两个连续的 Transformer 块中应用基于窗口的多头自注意力（W-MSA）模块和基于平移窗口的多头自注意力（SW-MSA）模块如图 2所示：

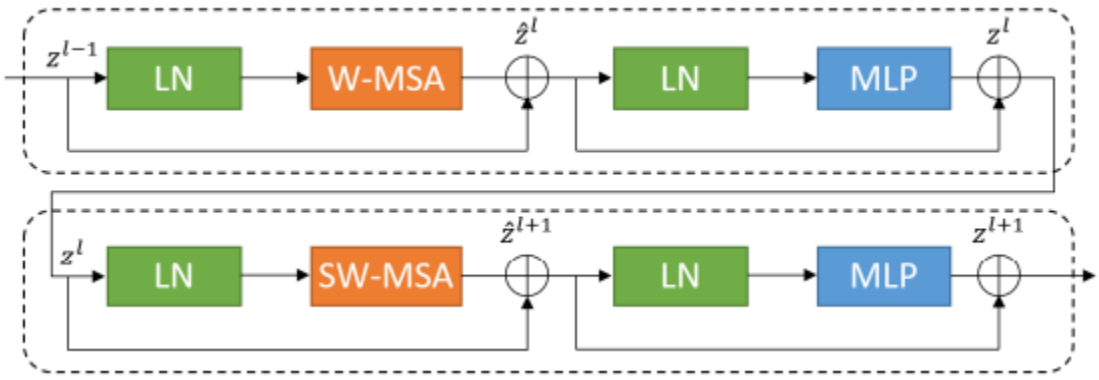

图 2: Swin transformer 模块图

3.3 补丁合并层

输入的图块被分成 4 个部分，并由补丁合并层连接在一起。通过这样的处理，特征分辨率将下采样 2 倍。而且，由于连接操作使特征维度增加了 4 倍，因此在连接的特征上应用线性层，将特征维度统一为原始维度的 2 倍。

3.4 补丁扩展层

以第一个图块扩展层为例，在上采样之前，对输入特征（W/32 × H/32 × 8C）应用线性层，将特征维度增加到原始维度的 2 倍（W/32 × H/32 × 16C）。然后，使用重新排列操作将输入特征的分辨率扩展到输入分辨率的 2 倍，并将特征维度减少到输入维度的四分之一（W/32 × /H32 × 16C → W/16 × H/16 × 4C）

4 复现细节

4.1 与已有开源代码对比

本次实验所使用的代码来自 swin-unet 的官方源码，主要的改动来自于补丁扩展层，将其替换为自己提出的一个双上采样模块，具体的结构创新点会详细介绍。

4.2 实验环境设置

本次实验所用的甲状腺结节数据集来源于深圳大学计算机与软件学院赖志辉教授团队，总共 982 张图像，训练集为 490 张，验证集为 492 张，由多位经验丰富的放射科医生提供分割结果。我们采用两个常用指标来定量评估不同分割模型的性能：并集交集（IoU）、Dice 系数。

4.3 复现细节

实验使用 Adamw 优化器来优化网络。初始学习率设置为 0.001，动量为 0.999。批量大小设置为 16，epoch 数为 300。此外，我们将所有图像的大小调整为 224×224，并执行随机旋转和翻转以进行数据增强。

4.4 创新点

对于上采样，原始的 Swin-UNet 采用 patch expanding 方法，等价于上采样模块中的转置卷积。然而，转置卷积很容易面对块效应。在这里，我们提出了一个新的模块，称为双上采样模块，它包括两种现有的上样本方法（即 Bilinear 和 PixelShuffle），以防止棋盘式的 artifacts。所提出的上采样模块的体系结构如下图所示。

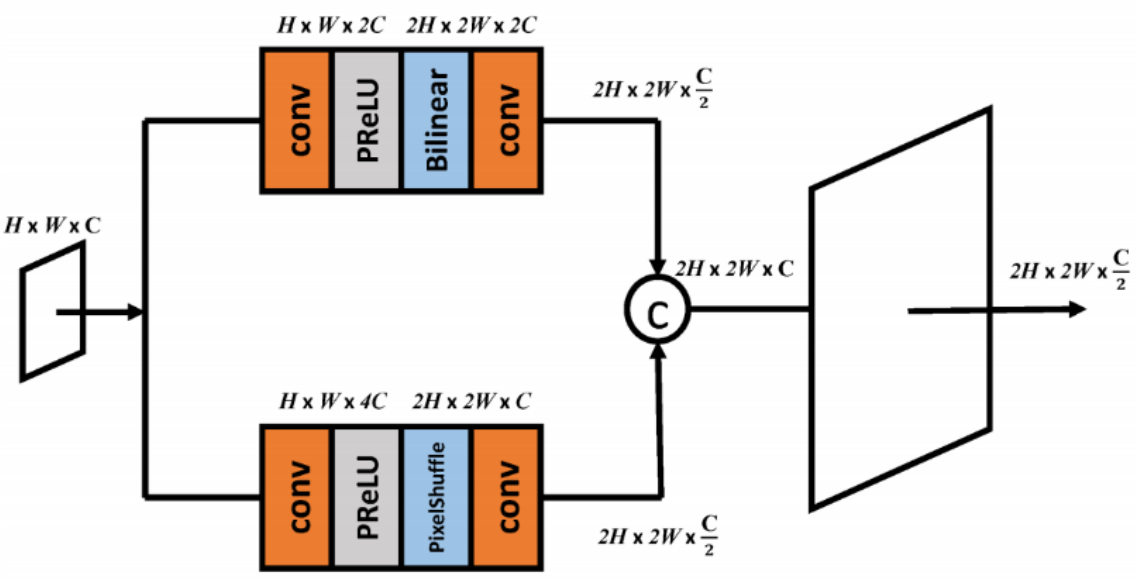

图 3: 双上采样模块图

5 实验结果分析

本次实验改进前后的性能对比可参考下图。总体而言，改进之后的 swin-unet 要比原始的 swin-unet 性能更好，在 IOU 和 Dice 系数，都提升了将近 1 个点。

	IOU	Dice
org swin-unet	72.52	82.75
improv swin-unet	73.18	83.47

图 4: 性能对比图

此外，训练损失的过程图如下图所示。可看到，改进之后的 swin-unet 总体收敛效果会更好，且收敛速度会更快。

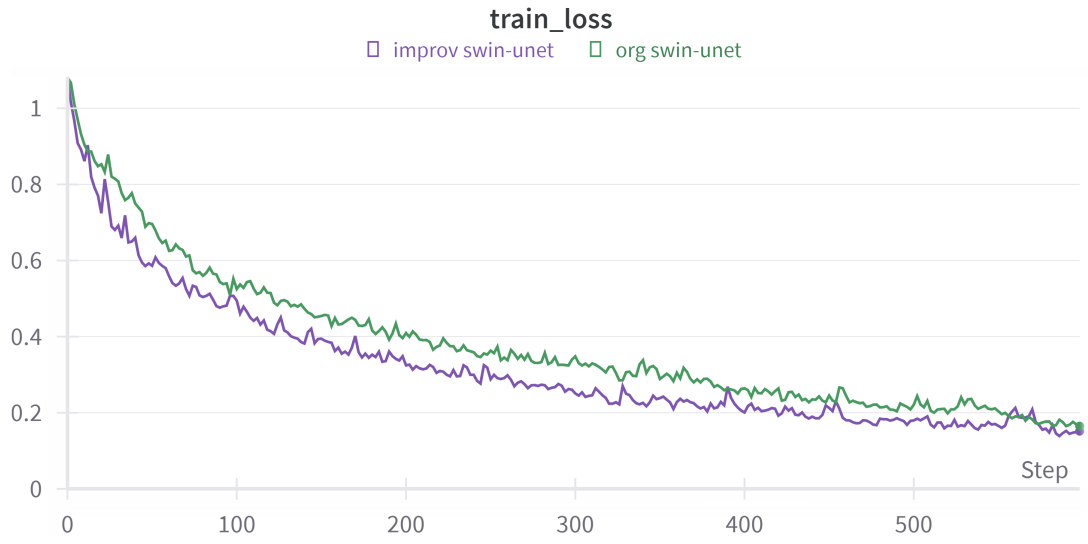

图 5: 训练损失图

可视化实验结果如下图所示。观察可得，改进之后的分割边缘会更好，且噪声点会更少，实现了更加精确的分割。

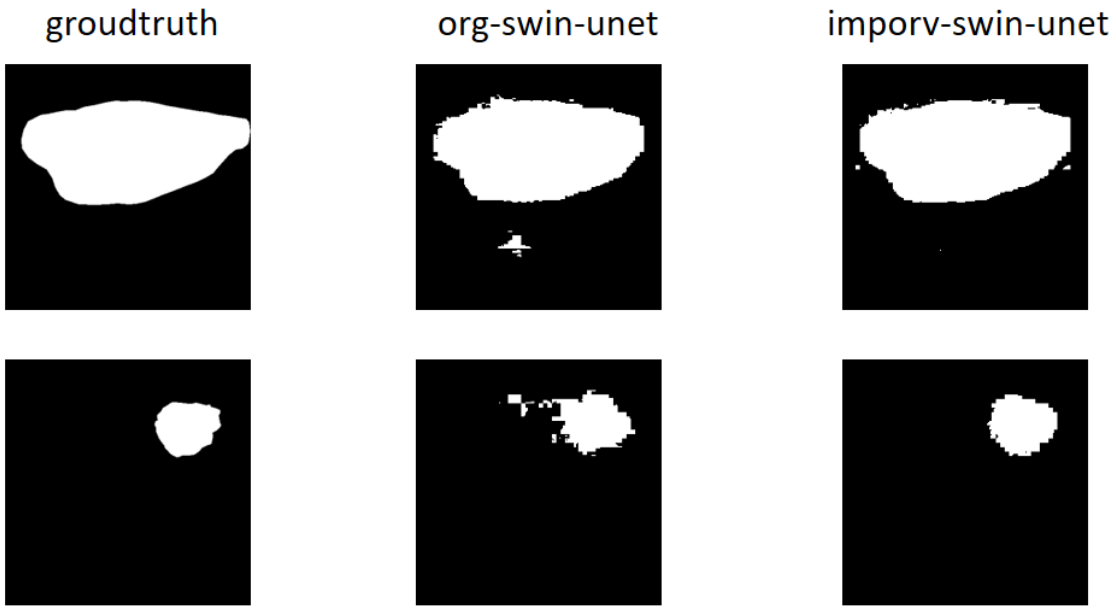

图 6: 可视化效果图

6 总结与展望

本次实验虽然对原始 swin-unet 进行了改进，但性能提升并不是特别多，个人认为可能是只改动补丁扩展层对于骨干网络提取特征并未产生太大影响。后续可以其他方面改进，例如改进自注意力机制，或者结合 CNN 来提取局部信息从而更有效地提取特征。此外，如何降低模型计算复杂度和参数量，也是一个很有意义地改进方向，因为模型落地部署往往需要的都是轻量级模型，只有模型落地部署了才能产生实际意义。

参考文献

[1] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. 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] JIN Q, MENG Z, SUN C, et al. RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans[J]. Frontiers in Bioengineering and Biotechnology, 2020, 8: 605132.

[4] ÇIÇEK Ö, ABDULKADIR A, LIENKAMP S S, et al. 3D U-Net: learning dense volumetric segmentation from sparse annotation[C]//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. 2016: 424-432.

[5] XIAO X, LIAN S, LUO Z, et al. Weighted res-unet for high-quality retina vessel segmentation[C]//2018 9th international conference on information technology in medicine and education (ITME). 2018: 327-331.

[6] ZHOU Z, RAHMAN SIDDIQUEE M M, TAJBAKHSH N, et al. Unet++: A nested u-net architecture for medical image segmentation[C]//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. 2018: 3-11.

[7] YIN X X, SUN L, FU Y, et al. U-Net-Based medical image segmentation[J]. Journal of Healthcare Engineering, 2022, 2022.

[8] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4): 834-848.

[9] GU Z, CHENG J, FU H, et al. Ce-net: Context encoder network for 2d medical image segmentation[J]. IEEE transactions on medical imaging, 2019, 38(10): 2281-2292.

[10] 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.

[11] 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.

[12] 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.

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[14] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]// European conference on computer vision. 2020: 213-229.

[15] 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.

[16] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention[C]//International conference on machine learning. 2021: 10347-10357.

[17] 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.

[18] TSAI A, YEZZI A, WELLS W, et al. A shape-based approach to the segmentation of medical imagery using level sets[J]. IEEE transactions on medical imaging, 2003, 22(2): 137-154.

[19] HELD K, KOPS E R, KRAUSE B J, et al. Markov random field segmentation of brain MR images[J]. IEEE transactions on medical imaging, 1997, 16(6): 878-886.

[20] LI X, CHEN H, QI X, et al. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes[J]. IEEE transactions on medical imaging, 2018, 37(12): 2663-2674.

[21] MILLETARI F, NAVAB N, AHMADI S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation[C]//2016 fourth international conference on 3D vision (3DV). 2016: 565-571.

[22] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[23] WANG W, XIE E, LI X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 568-578.

[24] HAN K, XIAO A, WU E, et al. Transformer in transformer[J]. Advances in Neural Information Processing Systems, 2021, 34: 15908-15919.