E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2023)11-3618-11

论文引用格式:Lu L and Qi W M. 2023. Spine CT image segmentation based on Transformer. Journal of Image and Graphics, 28(11):3618-3628(卢玲, 漆为民. 2023. 基于Transformer的脊椎CT图像分割. 中国图象图形学报, 28(11):3618-3628)[DOI:10.11834/jig.221084]

基于Transformer的脊椎CT图像分割

卢玲,漆为民*

江汉大学人工智能学院,武汉　430056

摘　要:**目的**　脊椎CT(computed tomography)图像存在组织结构显示不佳、对比度差以及噪音干扰等问题;传统分割算法分割精度低,分割过程需人工干预,往往只能实现半自动分割,不能满足实时分割需求。基于卷积神经网络(convolutional neural network,CNN)的U-Net模型成为医学图像分割标准,但仍存在长距离交互受限的问题。Transformer集成全局自注意力机制,可捕获长距离的特征依赖,在计算机视觉领域表现出巨大优势。本文提出一种CNN与Transformer混合分割模型TransAGUNet(Transformer attention gate U-Net),以实现对脊椎CT图像的高效自动化分割。**方法**　提出的模型将Transformer、注意力门控机制(attention gate,AG)及U-Net相结合构成编码—解码结构。编码器使用Transformer和CNN混合架构,提取局部及全局特征;解码器使用CNN架构,在跳跃连接部分融入AG,将下采样特征图对应的注意力图(attention map)与下一层上采样后获得的特征图进行拼接,融合低层与高层特征从而实现更精细的分割。实验使用Dice Loss与带权重的交叉熵之和作为损失函数,以解决正负样本分布不均的问题。**结果**　将提出的算法在VerSe2020数据集上进行测试,Dice系数较主流的CNN分割模型U-Net、Attention U-Net、U-Net++和U-Net3+分别提升了4.47%、2.09%、2.44%和2.23%,相较优秀的Transformer与CNN混合分割模型TransUNet和TransNorm分别提升了2.25%和1.08%。**结论**　本文算法较以上6种分割模型在脊椎CT图像的分割性能最优,有效地提升了脊椎CT图像的分割精度,分割实时性较好。

关键词:脊椎CT图像;医学图像分割;深度学习;Transformer;注意力门控机制(AG)

Spine CT image segmentation based on Transformer

Lu Ling, Qi Weimin*

School of Artificial Intelligence, Jianghan University, Wuhan　430056, China

Abstract: **Objective** The incidence of spine diseases has increased in the contemporary era and is increasingly affecting younger individuals. Therefore, the diagnosis and treatment of such diseases are particularly critical. Using 3D reconstruction technology, computer-aided diagnosis, and segmentation of the spine area and the background area of the spine computed tomography (CT) image can assist physicians in clearly observing the spine lesion area and provide theoretical support for surgical path simulation and surgical planning. The accuracy of spine CT image segmentation is critical in restoring the actual position and physiological shape of the patients' vertebrae to the greatest extent possible, thus allowing physicians to understand the distribution of lesions. However, the difficulty of spine segmentation is exacerbated by the complex structure of the spine, poor display of tissue structure, poor contrast, and noise interference in spine CT images. The segmentation of spine images via manual annotation relies on the physicians' a priori knowledge and clinical experience, and

收稿日期:2022-11-23;**修回日期**:2023-02-24;**预印本日期**:2023-03-03

*通信作者:漆为民　qwmin@jhun.edu.cn

基金项目:湖北省自然科学基金项目(2021CFB564)

Supported by: Natural Science Foundation of Hubei Province, China (2021CFB564)

the segmentation results are highly subjective and time consuming. Long working hours may also lead to deviations that affect the patients' diagnosis. With the help of computer technology, the traditional segmentation method mainly uses low-latitude features, such as texture, shape, and color of the image, for segmentation and often can only achieve semi-automatic segmentation. Moreover, this method does not fully utilize the image information and has low segmentation accuracy that fails to meet the demand of real-time segmentation. The segmentation method based on deep learning can realize automatic segmentation, effectively extract image features, and improve segmentation accuracy. In the branch of computer vision (CV), medical image segmentation algorithms based on convolutional neural network (CNN) have been proposed one after another and have become the mainstream research direction in medical image analysis. Among these algorithms, the characteristics of the U-Net structure itself and the fixed structure of medical images with multi-modality enhance the performance of U-Net in medical image segmentation and provide a benchmark for medical image segmentation. However, the inherent limitations of the convolutional structure can lead to problems, such as limited long-distance interaction. By contrast, Transformer, a non-CNN architecture, integrates a global self-attentive mechanism to capture long-range feature dependencies and is widely used in natural language processing, such as machine translation and text classification. In recent years, researchers have introduced Transformer into the field of computer vision and achieved advanced results in certain tasks, such as image classification and image segmentation. This paper then combines the advantages of the CNN architecture and Transformer to propose a CNN and Transformer hybrid segmentation model called Transformer attention gate U-Net (TransAGUNet) that realizes an efficient and automated segmentation of spine CT images. **Method** The proposed model combines Transformer, U-Net, and the attention gate (AG) mechanism to form an encoding-decoding structure. The encoder uses a hybrid Transformer and CNN architecture, which consists of a combination of ResNet50 and ViT models. For the sliced spine CT images, the low-level features are initially extracted by ResNet50, the feature maps corresponding to three downsampled features are retained, and then patch embedding and position embedding are performed. The obtained patches are then inputted to the Transformer encoder to learn long-term contextual dependencies and extract global features. The decoder adopts a CNN architecture that applies 2D bilinear upsampling at 2× rate to recover the image size layer by layer. The AG structure is incorporated into a jump-connected bottom-up triple layer to fuse shallow features with higher-level features for fine segmentation. The decoder uses a CNN structure to recover the image size layer by layer by performing 2D bilinear upsampling at a 2-fold rate. The AG structure is incorporated into the bottom-up three layers of the jump connection to obtain the attention map corresponding to the downsampled features, stitched with the upsampled features in the next layer, and then decoded by two ordinary convolutions and one 1 × 1 convolution. The AG structure then enters the binary classifier and distinguishes the foreground and background pixel by pixel to obtain the spine segmentation prediction map. The AG parameters are computationally small, easily integrated into CNN models, and can automatically learn the shape and size of the target to highlight salient features and suppress feature responses in irrelevant regions. These parameters replace the localization module via probability-based soft attention, thus eliminating the need to divide the ROI, and improve the sensitivity and accuracy of the model by a small amount of computation. The experiments use Dice Loss summed with weighted cross entropy loss as the loss function to solve the uneven distribution of positive and negative samples. **Result** The proposed algorithm is tested on the VerSe2020 dataset, and the Dice coefficients improve by 4.47%, 2.09%, 2.44%, and 2.23% over the mainstream CNN architectures of segmentation networks U-Net, Attention U-Net, U-Net++, and U-Net3+, respectively. Meanwhile, the Dice coefficients over the excellent Transformer and CNN hybrid segmentations TransUNet and TransNorm improve by 2.25% and 1.08%, respectively. To verify the validity of the proposed model, several ablation experiments are performed, and results show that compared with TransUNet, the Dice coefficient of the designed decoding structure improves by 0.75% and by 1.5% after adding AG. To explore the effect of the number of AG connections on the model performance, experiments are conducted using AG with different numbers of connections, and results show that the Dice coefficient obtained without adding AG is the smallest and that the optimal model performance is achieved by adding AG in three jump connections on the resolution scales of 1/2, 1/4, and 1/8. **Conclusion** Compared with the above six CNN segmentation models and the Transformer and CNN hybrid segmentation models, the proposed algorithm achieves the best segmentation results on spine CT images, thus effectively improving the segmentation accuracy of spine CT images with better segmentation real-time performance.

Key words: spine CT image; medical image segmentation; deep learning; Transformer; attention gate (AG)

0 引 言

脊椎疾病已成为当代高发疾病且呈年轻化发展趋势,因此其诊断和治疗尤为关键。随着人工智能技术的不断发展,智能化诊断的需求不断提高。在计算机辅助诊断下,使用分割算法分割出脊椎CT(computed tomography)影像中的感兴趣区域,结合三维重建技术,可使医生直观清晰地观察和剖析病灶区域,为模拟手术路径及外科手术方案制定提供理论支撑,提高诊断效率和正确率。然而由于脊椎结构复杂,脊椎CT影像中存在噪音干扰,脊椎边缘模糊,分界不清等问题,加剧了脊椎分割的难度。

随着人工智能技术和深度学习方法的迅猛发展,基于卷积神经网络(convolutional neural network, CNN)的深度学习算法在医学图像分割领域上取得了显著成效。Long等人(2015)提出了首个端对端的针对像素级预测的全卷积神经网络(full convolutional networks, FCN),解决了语义级别的图像分割问题。但FCN对细节信息不敏感,分割不够精细,针对此问题,Ronneberger等人(2015)提出了一种基于FCN的U-Net图像分割模型,采用编码器—解码器结构及跳跃连接的设计模式,将浅层特征和深层特征进行融合,实现更精细的分割。由于U-Net结构自身特点及医学图像结构固定、具有多模态等特点,使得U-Net在医学图像分割上表现良好,成为医学图像分割的基准。由于卷积运算固有的局限性,导致基于CNN的分割模型如U-Net存在长距离交互受限等问题。Transformer(Vaswani等,2017)集成全局自注意力机制,可捕获长距离的特征依赖,在自然语言处理(natural language processing, NLP)取得了广泛的成功。Dosovitskiy等人(2021)基于此提出了ViT(vision Transformer)模型,在图像识别任务中获得了更高的性能。此后,Transformer便更广泛地运用到计算机视觉领域,并表现出巨大的优势。由于Transformer的计算量很大且不能有效地捕获区域特征,考虑到CNN获取局部特征及Transformer捕获全局特征的优势,许多研究人员将U-Net和Transformer进行结合并应用到语义分割任务(Chen等,2021; Guo和Terzopoulos, 2021; Azad等,2022)中,以捕获局部和全局特征,获得更好的分割性能。

基于深度学习的分割方法已在脊椎图像的分割问题上有了成功应用。刘忠利等人(2018)基于FCN提出卷积、反卷积神经网络模型对椎骨进行全自动分割。李贤和何洁(2018)使用3D全卷积网络分割椎骨,缩短了分割时间,分割效果较好。Kolařík等人(2019)使用3D Dense U-Net分割胸椎和腰椎。田丰源等人(2020)使用AttentionNet(Sekuboyina等,2017)定位脊椎,再使用改进的Dense-UNet(Li等,2018))分割脊椎,分割精度优于传统Dense-UNet。金顺楠等人(2021)将尺度残差模块及通道注意模块引入到U-Net网络中分割脊椎椎骨,获得了较高的分割精度及分割效率。基于深度学习的方法在医学图像分割上效果显著,但对分割精度有了更高的要求。

由于大多数对脊椎图像的研究工作是基于CNN模型开展的,存在一定的局限性,分割精度还有待提升。Transformer近年来才被应用到计算机视觉领域中且取得了一定的成功,但在脊椎CT图像的分割任务上研究甚少。故本文以脊椎CT图像作为研究对象,旨在提出一种基于Transformer的分割算法,结合CNN与Transformer的优势,实现对脊椎CT图像的高效自动化分割,提高脊椎分割精度。主要研究内容包括:1)结合Transformer、注意力门控机制(attention gate, AG)(Oktay等,2018)及U-Net网络,提出一种CNN与Transformer的混合分割模型TransAGUNet(Transformer attention gate U-Net),实现脊椎CT图像的自动化分割,以解决U-Net远距离传输受限、Transformer局部特征识别不足等问题,进一步提升脊椎分割精度。TransAGUNet为编码器—解码器结构,编码器采用CNN和Transformer混合架构,获取丰富的局部与全局信息,其结构类似于TransUNet的编码结构。解码器由CNN架构组成,在跳跃连接中融入AG,将得到的注意力图与解码器上采样获得的特征图进行拼接,融合低层与高层特征从而实现更精细的分割。将经过跳跃连接后拼接的特征图进行两次卷积操作,增强网络对脊椎的特征提取能力,再进行一次1 × 1卷积降维,减少网络参数量;2)设计对比实验和消融实验,验证模型的有效性。实验使用Dice Loss与带权重的交叉熵之

和作为损失函数，以解决正负样本分布不均的问题。将提出的模型在 VerSe2020（vertebrae segmentation）数据集上测试，分割结果在其余 6 种 CNN 分割模型及 Transformer 与 CNN 混合分割模型中最佳。

1 基于深度学习的医学图像分割模型

1.1 CNN 分割模型

基于卷积神经网络（CNN）的分割模型已成功地应用在众多医学图像分割任务中，如脑肿瘤分割（赵奕名 等，2020）、胸部多器官分割（吉淑滢和肖志勇，2021）、淋巴结分割（刘羽 等，2022）等。由于 U-Net（Ronneberger 等，2015）在医学图像分割上取得了很好的效果，一系列 U-Net 的变型网络模型被相继提出。Oktay 等人（2018）提出 Attention U-Net，将提出的注意力门控机制（AG）与 U-Net 相结合，首次在医学图像的 CNN 中使用 soft attention，增加了模型对前景像素的敏感度，基于网格的 AG 使注意力系数更关注局部区域特征，抑制无关区域。之后，Xiao 等人（2018）针对视网膜血管成像限制及光源干扰等分割任务的难点提出了 Res-UNet（residual U-Net），该模型将残差网络 ResNet（residual neural network）（He 等，2016）和 U-Net 进行了融合，增加了网络的深度，防止过拟合，提高了模型的准确度。Zhou 等人（2018）基于 DenseNet（dense network）（Huang 等，2017）思想提出 U-Net++，使用密集的跳跃连接，通过特征叠加的方式整合不同的特征，是一种深度监督的编码器—解码器网络。Jha 等人（2019）提出 Res-UNet++，在 Res-UNet 的基础上对图像的后处理部分使用了条件随机场（conditional random filed，CRF）及测试时数据增强（test time augmentation，TTA），使用空洞空间卷积池化金字塔模块（atrous spatial pyramid pooling，ASPP）代替 Res-UNet 中的桥接部分，分割性能优于 Res-UNet，在难以分辨的息肉问题上表现优异。Huang 等人（2020）提出 U-Net3+，表示 U-Net++ 虽然使用了密集的跳跃连接，但未充分利用多尺度提取足够信息，因此在 U-Net3+ 中提出了全尺度跳跃连接（full-scale skip connections），精度较 U-Net++ 有一定的提升。虽然这些方法可在一定程度上提高医学图像的分割精度，但仍存在长距离交互受限、全局信息提取不足等问题。

1.2 Transformer 分割模型

Transformer（Vaswani 等，2017）最初应用于自然语言处理并在很多任务中获得了巨大的成功，如释义短语生成（Egonmwan 和 Chali，2019）、语音识别（Shi 等，2021）及语音合成（Chen 和 Rudnicky，2022）等。受此启发，研究人员将其运用到计算机视觉领域，在图像分类（Dosovitskiy 等，2021）、语义分割（Strudel 等，2021）等计算机视觉（computer vision，CV）任务中应用广泛。Dosovitskiy 等（2021）提出的 ViT 模型较传统的 CNN 有更高的性能。ViT 模型将输入图像分成固定大小的图像块 Patches，然后通过线性变换得到 Patch embedding，并使用 Position embedding 编码位置信息，再将经过以上处理的 Patches 输入到 Transformer 的编码器中进行特征提取，最后通过多层感知机（multi-layer perceptron，MLP）完成分类。Segmenter（Transformer for semantic segmentation）（Strudel 等，2021）是基于 ViT 改进的纯 Transformer 图像语义分割模型，是一种完全基于 Transformer 的编码器—解码器架构，编码器采用 ViT 类似结构，解码器使用逐点线性映射或 mask Transformer，可以很好地捕获全局上下文信息，提高了图像分割性能。

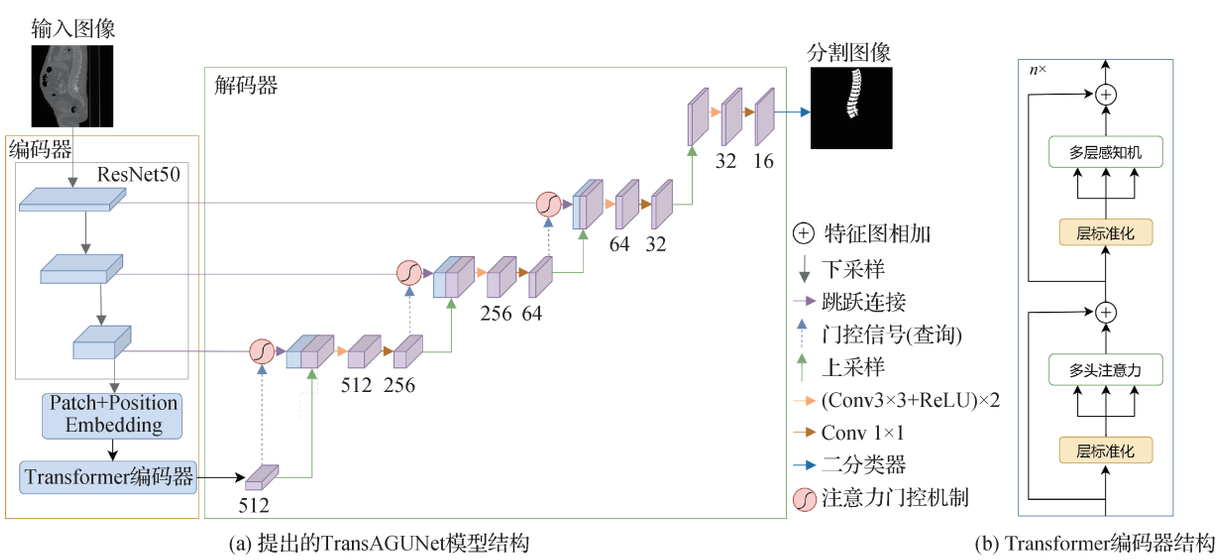
1.3 CNN 与 Transfomer 混合分割模型

对于医学图像分割问题，大部分的研究工作是基于 CNN 分割模型展开的，近年来由于 Transformer 在计算机视觉领域取得了重大的突破，研究者将 CNN 与 Transformer 相结合，提出的混合分割模型较仅有的 CNN 模型更具全局特征提取能力，在医学图像分割问题上取得了进一步成功。Chen 等人（2021）提出的 TransUNet 分割模型首先利用 CNN（ResNet50）提取低级特征，然后使用 ViT 进行编码，对全局交互进行建模，并结合跳跃连接，在 Synapse 多器官分割数据集上分割性能优于 U-Net、Attention U-Net，成为医学图像分割的强大替代方案。随后，一系列基于 CNN 与 Transformer 的混合分割模型相继提出，如 TransBTS（Transformer brain tumor segmentation）（Wang 等，2021）、nnFormer（not-another Transformer）（Zhou 等，2022）、TransNorm（Transformer spatial normalization）（Azad 等，2022）等。TransBTS 首次使用 3D CNN 中的 Transformer 分割 MRI（magnetic resonance imaging）脑肿瘤，编码器首先使用 3D CNN 提取空间特征图，然后将特征图映

射并改进后的tokens传入Transformer中进行全局建模，解码器采用渐进式上采样得到预测的分割图，在BraTS2019（brain tumor segmentation）数据集上进行测试，分割性能优于最先进的3D MRI脑肿瘤分割方法。TransNorm从Transformer模块中推导出一个空间归一化模块，与跳跃连接后的特征图进行拼接，自适应校准跳跃连接路径，在Synapse、ISIC2017（international skin imaging collaboration）、ISIC2018这3个经典的医学图像分割数据集上均取得了较好的分割性能，分割精度高于TransUNet。

2 本文方法

针对脊椎CT图像分割，本文结合Transformer、AG和U-Net，提出一种CNN与Transformer混合分割模型TransAGUNet（Transformer attention gate U-Net），构成编码—解码结构，其模型结构如图1所示。编码结构采用CNN与Transfomer混合架构，具体由ResNet50与ViT模型（Dosovitskiy，2021）组合构成，与TransUNet（Chen等，2021）的编码结构类似。对于输入的脊椎图像，首先通过ResNet50提取低级特征，保留3次下采样对应的特征图，然后进行块编码（Patch embedding）与位置编码（Position embedding），将得到的patches输入到Transformer编码器中，学习长期上下文依赖关系，提取全局特征。在解码部分前3层的跳跃连接中融入AG，得到下采样特征图对应的注意力图（attention map），再与下一层经过上采样后的特征图进行拼接，然后进行两次普通卷积及一次1×1卷积进行解码。最后一层将上一层上采样后的特征图通过两次普通卷积与一次1×1卷积，最后进入二分类器，逐像素区分前景和背景，得到脊椎分割预测图。



(a) 提出的TransAGUNet模型结构　(b) Transformer编码器结构

图1 模型总体结构

Fig. 1 General structure diagram of the model

((a) structure diagram of the proposed TransAGUNet model; (b) structure diagram of Transformer encoder)

2.1 编码器

2.1.1 Embedding

模型编码器中的Embedding部分包括Patch embedding和Position embedding。用H、W表示图像的高、宽，C表示图像通道数。对于输入维度为$H \times W \times C$的图像，Patch embedding操作将图像重塑（reshape）成维度为$N \times P^2 \times C$的patches $\boldsymbol{x}_{\mathrm{p}}$。其中，$N = HW/P^2$，每个patch的大小为$P \times P$，通道数为$C$，分别用$x_{\mathrm{p}}^1, x_{\mathrm{p}}^2, \cdots, x_{\mathrm{p}}^N$表示，然后使用线性投影$\boldsymbol{E}$将patches映射到$D$维空间。为了保留patches的空间信息，对其叠加Position embedding，用$\boldsymbol{E}_{\mathrm{pos}}$表示。整个过程可表示为

$$z_0 = \left[\boldsymbol{x}_{\mathrm{p}}^1\boldsymbol{E}; \boldsymbol{x}_{\mathrm{p}}^2\boldsymbol{E}; \ldots \boldsymbol{x}_{\mathrm{p}}^N\boldsymbol{E}\right] + \boldsymbol{E}_{\mathrm{pos}} \tag{1}$$

式中，z_0表示经过Embedding层后得到的特征图，$\boldsymbol{E} \in \mathbf{R}^{P^2 \times C \times D}$，$\boldsymbol{E}_{\mathrm{pos}} \in \mathbf{R}^{N \times D}$。

2.1.2 Transformer编码器

Transformer编码器将图像经过Embedding得到的patches作为输入，其整体结构如图1（b）所示，由n

层构成，本文采用的 n 为 12。每一层由多头自注意力（multi-head self-attention，MSA）和多层感知机（MLP）模块组成。其中 MLP 由两层线性层组成，两层均使用 GELU（Gaussian error linear unit）作为激活函数。Transformer encoder 第 n 层的输出可表示为

$$z'_n = MSA(LN(z_{n-1})) + z_{n-1} \tag{2}$$

$$z_n = MLP(LN(z'_n)) + z'_n \tag{3}$$

式中，LN 代表层标准化操作（layer normalization）（Ba 等，2016），z_n 表示经过编码后的图像表示。

2.2　解码器

2.2.1　注意力门控机制

CNN 在对形变程度较大的医学图像进行分割时，通常采取的做法是先定位，确定感兴趣区域（region of interest，ROI），再进行分割。注意力门控机制 AG 参数计算量小，很容易与 CNN 模型进行整合，将 CNN 与 AG 进行结合，也可达到此效果。AG 自动学习目标的外形和尺寸，突出显著特征，抑制无关区域的特征响应，通过基于概率的 soft attention 替代定位模块，无需划分 ROI，通过少量计算量来提高模型的敏感度与准确率。AG 的具体结构图如图 2 所示。

首先将经过上采样后维度为 $H \times W \times C$ 的特征图 $\boldsymbol{x}_u$ 与经过 CNN 提取的维度为 $H \times W \times C$ 的特征图 $\boldsymbol{x}_s$ 进行并行处理，分别使用 3 × 3 的卷积及批归一化（batch normalization，BN）操作得到维度为 $H \times W \times (C/2)$ 的 $\boldsymbol{x}'_u$ 和 $\boldsymbol{x}'_s$，再将 $\boldsymbol{x}'_u$ 与 $\boldsymbol{x}'_s$ 对应元素相加，然后进行 ReLU（rectified linear unit）操作，随后进行 1 × 1 输出通道数为 1 的卷积操作，再使用 BN、sigmoid 激活函数得到维度为 $H \times W \times 1$ 的注意力系数权重 $\boldsymbol{\alpha}$，最后使用 $\boldsymbol{x}_s$ 乘以 $\boldsymbol{\alpha}$，得到维度为 $H \times W \times C$ 的注意力特征图 $\boldsymbol{x}_r$。

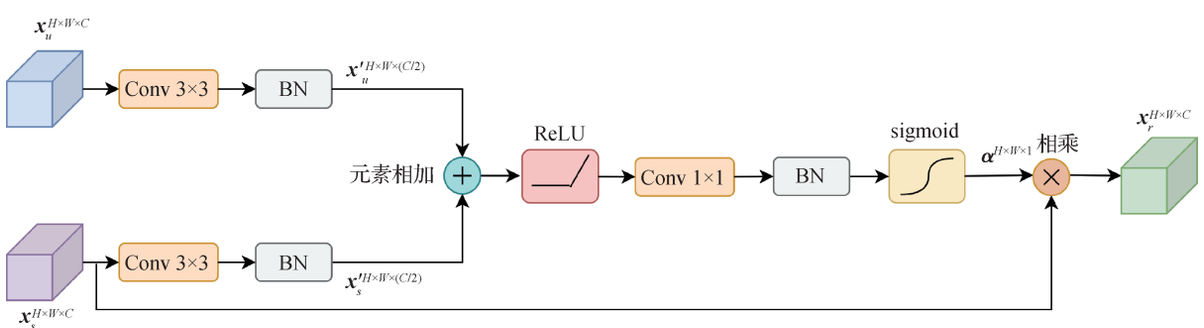


图 2　AG 结构图

Fig. 2　AG structure diagram

2.2.2　解码结构

解码器采用的是 CNN 架构，使用二维双线性上采样 2 倍率逐层恢复图像尺寸。在跳跃连接自下而上的 3 层中融入 AG 结构，将浅层特征与高层特征进行融合，实现精细分割。以第 3 层结构为例，首先将 CNN 提取的特征图 $\boldsymbol{x}_s \in \mathbf{R}^{H' \times W' \times C'}$ 经过 AG 得到注意力特征图 $\boldsymbol{x}_r$，再将 $\boldsymbol{x}_r$ 与下一层经过上采样后的特征图 $\boldsymbol{x}_u$ 在通道维度上进行拼接，得到特征图 $\boldsymbol{x}_1 \in \mathbf{R}^{H' \times W' \times 2C'}$；然后进行两次 3 × 3 输出通道数为 C' 的卷积操作，使用 ReLU 激活函数，得到特征图 $\boldsymbol{x}_2 \in \mathbf{R}^{H' \times W' \times C'}$；再使用 1 × 1 输出通道数为 C'' 的卷积操作降维，得到特征图 $\boldsymbol{x}_3 \in \mathbf{R}^{H' \times W' \times C''}$。本文 4 次上采样后设置的输出通道数分别为［256，64，32，16］，最后一次上采样特征图经过两次 3 × 3 及一次 1 × 1 卷积操作后，得到特征图 $x \in \mathbf{R}^{2H' \times 2W' \times 16}$，解码器具体结构如图 1（a）所示。

2.3　损失函数

2.3.1　带权重的交叉熵损失函数

交叉熵损失（cross entropy loss，CE Loss）是基于分布的损失函数，网络训练过程中梯度下降更新快，常作为分类器的损失函数，对每个类别的权重相同，计算式为

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=0}^{m-1} y_{ij}\log(p_{ij}) \tag{4}$$

式中，N 表示样本个数，m 表示样本分类数，y_{ij} 表示真实值，p_{ij} 表示预测值。

对于医学图像分割任务，往往是对 CT 切片后的图像进行逐像素分类，划分前景和背景区域。通常伴随前景和背景分布不均的问题，即背景像素偏多，前景像素偏少的情况，导致模型训练更易于学习背景特征，而很难学习前景特征，从而降低模型对前景区域的分割精度。因此对 CE Loss 进行改进，从而得

到带权重的交叉熵损失(weighted cross entropy loss, WCE Loss),对较少类别进行加权,计算式为

$$L_{\mathrm{WCE}}=-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=0}^{m-1}w_j y_{ij}\log(p_{ij}) \tag{5}$$

式中,w_j为每个类别的权重。

2.3.2 Dice Loss

Dice Loss由Milletari等人(2016)为应对语义分割任务中正负样本不平衡问题而提出。来源于用来评估样本相似度的度量函数Dice相似系数,计算式为

$$L_{\mathrm{Dice}}=1-\frac{2|\boldsymbol{X}\cap\boldsymbol{Y}|}{|\boldsymbol{X}|+|\boldsymbol{Y}|} \tag{6}$$

式中,$\boldsymbol{X}$和$\boldsymbol{Y}$分别表示真实和预测轮廓区域所包含的点集。

Dice Loss是一个区域相关的损失函数,即当前像素点的损失及梯度值与该点及其他像素点的预测值及真实结果(ground truth)相关。Dice Loss对于固定大小的正样本区域计算的损失是相同的,且在训练过程中更倾向于挖掘前景区域,从而在一定程度上解决正负样本不均的问题。

本文主要分割出脊椎与非脊椎部分,即逐像素区分前景和背景两类。为解决脊椎图像前景和背景像素不平衡问题,且考虑到Dice Loss训练不稳定,在极端情况下会出现梯度饱和现象,因此结合WCE Loss进行改进,在模型训练中采用Dice Loss + WCE Loss作为损失函数,从而提高模型的分割精度。

3 实 验

3.1 数据集及预处理

实验使用的CT数据集来自于国际医学图像计算和计算机辅助干预协会(Medical Image Computing and Computer Assisted Intervention Society, MICCAI) 2020年举办的脊椎分割挑战赛数据集VerSe2020(Löffler等,2020)。VerSe2020包含训练集和测试集各100例,包括颈椎(C1-C7)、胸椎(T1-T12)和腰椎(L1-L5),是目前为止最大的脊椎分割数据集,其分割真实图像由专业医生手工标注。从100例数据集中分别筛选出72例、8例作为本次实验的训练集、验证集,然后在测试集中随机选取16例作为本次实验的测试集。

首先将选取的CT数据统一调整为RAI(right anterior inferior)方向,再进行切片处理。由于相邻片非常相似,防止产生过多冗余数据,本次实验采取间隔2片进行切片操作,将3D体素数据转化为2D图像,并舍弃只含有背景部分的图像,最终得到训练集图像3 168幅,验证集图像403幅,测试集图像774幅。为了减少网络训练计算量,将所有图像均裁剪为256×256像素,并转化为对应数据集的npy(numpy)文件,以提高数据读取速度。

3.2 实验环境

实验基于Ubuntu16.04操作系统,使用4块显存为8 GB的NVIDIA GeForce GTX 1070Ti显卡,分布式数据并行(distributed data parallel, DDL)模式进行多卡并行训练,使用Python3.8作为开发语言,开发框架为PyTorch1.11。实验batch size设为8,epochs为100,使用梯度下降法(stochastic gradient descent, SGD)优化器,初始学习率设为0.04、动量为0.9、权重衰减率为0.000 1。学习率采用动态更新策略,在每一次迭代中根据学习轮次线性降低。使用正态分布对数据进行初始化,在编码和解码阶段加入BN层以加速网络收敛。

3.3 评价指标

本文主要采用Dice相似系数(Dice similarity coefficient, DSC)、均交并比(mean intersection over union, mIoU)、召回率(recall)、精确率(precision)和像素准确率(pixel accuracy, PA)作为评价指标来评估模型对脊椎CT图像的分割性能。其中,使用mIoU和PA作为评价指标,不仅考虑到对脊椎的分割,同时考虑到对背景的精确识别,通过混淆矩阵(confusion matrix)来实现。

3.4 对比实验

为客观评估提出方法的分割性能,在相同实验环境及数据集下,将提出的CNN与Transformer混合分割模型TransAGUNet与优秀的CNN分割模型U-Net、Attention U-Net、U-Net++、U-Net3+及CNN与Transformer混合分割模型TransUNet、TransNorm的测试结果进行对比,实验结果如表1所示。

由表1可见,U-Net模型脊椎分割的Dice系数为0.743 1,其余各模型较此均有一定的提升。其中,Attention U-Net在CNN架构的对比分割模型中Dice系数提升最为显著,提升了2.38%,可见Attention U-Net在U-Net的基础上融入AG后,模型的性能得到了显著地提升,因而本文在模型设计过程中考虑

到了 AG 的融入。提出的 TransAGUNet 模型在 Dice 系数、mIoU 及召回率这 3 个评价指标上都取得了最好的结果。TransAGUNet所得的Dice系数为0. 787 8，较CNN 分割模型 U-Net、Attention U-Net、U-Net++、U-Net3+分别提升了4. 47%、2. 09%、2. 44%，2. 23%，较 Transformer 与 CNN 混合分割模型 TransUNet、TransNorm 分别提升了 2. 25%、1. 08%；mIoU 达到了 0. 895 2，与以上 6 种分割模型相比，分别提高了 0. 78%、0. 42%、0. 06%、0. 32%、0. 93% 和 1. 14%，反映了 TransAGUNet 分割结果与真实值的高相似度；召回率达到了 0. 848 7，表明模型能较准确地识别前景部分；PA 达到了 0. 994 0，由于背景部分占比较大，PA 值主要反映了对模型对背景的精确识别能力。综上可知，本文提出的模型分割结果与真实值相似度较高，能较好地识别前景与背景，分割性能优于其余 6 种模型，有效地提升了脊椎 CT 图像的分割精度。

表 1　不同模型在 VerSe2020 数据集上的分割结果对比

Table 1　Comparison of segmentation results of different models on the VerSe2020 dataset

模型	Dice 相似系数	均交并比	召回率	精确率	像素准确率
U-Net	0.743 1	0.887 4	0.816 3	0.780 2	0.993 7
Attention U-Net	0.766 9	0.891 0	0.815 8	0.800 4	0.993 9
U-Net++	0.763 4	0.894 6	0.818 7	0.800 2	**0.994 1**
U-Net3+	0.765 5	0.892 0	0.804 5	**0.839 7**	0.994 0
TransUNet	0.765 3	0.885 9	0.825 0	0.782 4	0.993 3
TransNorm	0.777 0	0.883 8	0.838 2	0.777 6	0.993 3
TransAGUNet (本文)	**0.787 8**	**0.895 2**	**0.848 7**	0.792 3	0.994 0

注：加粗字体表示各列最优结果。

3. 5　消融实验

3. 5. 1　提出的解码结构及 AG 对模型性能的影响

TransUNet 是目前较新的、分割性能较优的 CNN 与 Transformer 混合分割模型，从表 1 可知，其 Dice 系数与 mIoU 均高于 U-Net。由于提出的模型与 TransUNet 的 Backbone 部分类似，较 TransUNet 相比，主要体现在解码结构及跳跃连接部分的不同，为了更好地反映提出的 CNN 解码结构及在跳跃连接部分融入 AG 对模型性能的影响，以 TransUNet 作为基准，设置相应的消融实验，实验结果如表 2 所示。其中，Model1 与 TransUNet 相比改变了解码结构，Model2 在 Model1 的基础上在跳跃连接部分加入了 AG，即本文提出的 TransAGUNet 模型。

从表 2 可见，TransUNet 的 Dice 系数为 0. 765 3，mIoU 为 0. 885 9，召回率为 0. 825 0。Model1 与 TransUNet 相比，性能有了一定的提升，Dice 系数、mIoU 及召回率分别提升了 0. 75%、0. 45% 和 2. 34%。由此可见，本文设计的解码结构可有效地融合编码结构提取的特征，恢复图像大小。Model2 在 Model1 的基础上 Dice 系数提升较大，提升了 1. 5%，可知加入 AG 后，增强了对显著特征的提取能力，分割性能得到了明显提升。

表 2　提出的解码结构及 AG 对模型性能的影响

Table 2　The influence of proposed decoding structure and AG on model performance

模型	Dice 相似系数	均交并比	召回率	精确率	像素准确率
TransUNet	0.765 3	0.885 9	0.825 0	0.782 4	0.993 3
Model1	0.772 8	0.890 4	0.848 4	0.770 7	0.993 7
Model2	**0.787 8**	**0.895 2**	**0.848 7**	**0.792 3**	**0.994 0**

注：加粗字体表示各列最优结果。

3. 5. 2　AG 连接数量对模型性能的影响

为了进一步探究 AG 在跳跃连接中的连接数量对模型性能的影响，在模型结构中的跳跃连接中使用不同数量的 AG 进行实验，实验结果如表 3 所示。其中，AG = 0 表示在跳跃连接中不加入 AG，即为上文中的 Model1，AG = 3 即为本文所提出的 TransAGUNet 模型。从表 3 可见，Dice 系数随跳跃连接中 AG 数量的增加而增加，不加入 AG 所得到的 Dice 系数最小，在 1/2、1/4、1/8 分辨率尺度上的 3 个跳跃连接中加入 AG，所得的模型性能最优。

表3　AG连接数量对模型性能的影响

Table 3　The Influence of the number of AG connections on the model performance

连接数量	Dice相似系数	均交并比	召回率	精确率	像素准确率
AG = 0	0.772 8	0.890 4	0.848 4	0.770 7	0.993 7
AG = 1	0.776 0	0.888 9	0.831 5	0.794 3	0.993 5
AG = 2	0.781 1	0.894 2	0.835 6	0.799 4	0.993 9
AG = 3	**0.787 8**	**0.895 2**	**0.848 7**	**0.792 3**	**0.994 0**

注：加粗字体表示各列最优结果。

3.6　分割结果可视化对比

为了将本文模型与TransUNet模型的分割结果进行更直观的展示与对比，在测试集上将两者的分割预测图均转化为灰度图，并选取部分数据，其结果同输入图像及标签的对比如图3所示。其中，第1幅测试图的分割部位为C1-C7和T1-T2；第2幅测试图的分割部位为T10-T12和L1-L5；第3幅测试图的分割部位为T2-T12和L1-L5。

从图3（c）中可见TransUNet在第1幅分割图中对脊椎分割细节上缺乏一定的敏感度，存在误分割现象，错将背景预测为颈椎椎骨；在第2幅分割图中存在欠分割现象，未能完整分割出最后一节的腰椎椎骨结构；在第3幅分割图上既错将背景预测为胸椎椎骨，又缺乏对腰椎的完整分割。图3（d）中即提出的模型中在跳跃连接加入了AG结构后，加强了对椎体结构的识别能力，语义信息丢失问题也得到了改进，欠分割问题大大减少，分割结果更接近标签值，分割性能更好。

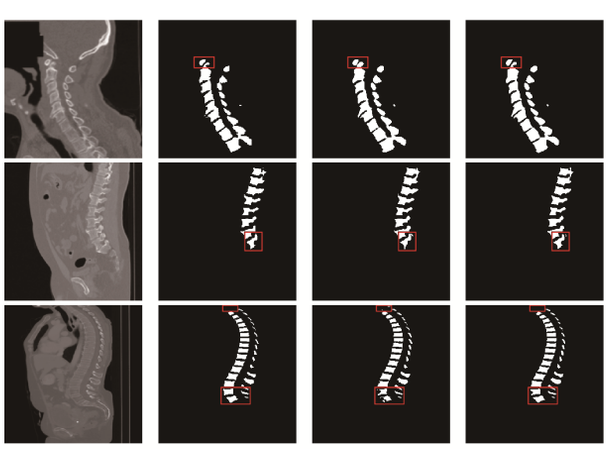


图3　分割结果可视化对比图

Fig. 3　Comparative visualization of segmentation results ((a) input images; (b) ground truth; (c) TransUNet; (d) ours)

4　结　论

本文结合Transformer、AG和U-Net，提出一种CNN与Transformer混合分割模型TransAGUNet，实现对脊椎CT图像的全自动分割。TransAGUNet使用Transformer和CNN混合架构作为编码器，提取语义和远程上下文特征；使用CNN结构作为解码器，在跳跃连接部分加入注意力门控机制AG，加强对显著目标区域的特征提取，抑制无关区域；使用Dice Loss与带权重的交叉熵之和作为损失函数以解决正负样本不均衡的问题。实验结果表明，本文模型与其余6种对比网络模型包括CNN分割模型及Transformer与CNN混合分割模型在脊椎CT图像的自动分割任务上取得了最高的分割精度，同时表明在Transformer与CNN混合分割模型中加入注意力门控机制能有效地提高脊椎CT图像的分割精度。本文算法对使用深度学习算法分割脊椎CT图像的研究工作提供了重要参考，但仍存在分割细节不足的问题，如对于部分腰椎结构未能完整分割出来，模型设计仍有改进的地方，这也是今后要研究的重点。

参考文献（References）

Azad R, Al-Antary M T, Heidari M and Merhof D. 2022. TransNorm: Transformer provides a strong spatial normalization mechanism for a deep segmentation model. IEEE Access, 10: 108205-108215 [DOI: 10.1109/ACCESS.2022.3211501]

Ba J L, Kiros J R and Hinton G E. 2016. Layer normalization [EB/OL]. [2022-11-23]. https://arxiv.org/pdf/1607.06450.pdf

Chen J N, Lu Y Y, Yu Q H, Luo X D, Adeli E, Wang Y, Lu L, Yuille A L and Zhou Y Y. 2021. TransUNet: Transformers make strong encoders for medical image segmentation [EB/OL]. [2022-11-23]. https://arxiv.org/pdf/2102.04306.pdf

Chen L W and Rudnicky A. 2022. Fine-grained style control in Transformer-based text-to-speech synthesis//Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore, Singapore: IEEE: 7907-7911 [DOI: 10.1109/ICASSP43922.2022.9747747]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16 × 16 words: Transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. [s.

l.]: OpenReview.net, 2021

Egonmwan E and Chali Y. 2019. Transformer and seq2seq model for paraphrase generation//Proceedings of the 3rd Workshop on Neural Generation and Translation. Hong Kong, China: Association for Computational Linguistics: 249-255 [DOI: 10.18653/v1/D19-5627]

Guo D F and Terzopoulos D. 2021. A Transformer-based network for anisotropic 3D medical image segmentation//Proceedings of the 25th International Conference on Pattern Recognition. Milan, Italy: IEEE: 8857-8861 [DOI: 10.1109/ICPR48806.2021.9411990]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/cvpr.2016.90]

Huang G, Liu Z, Van Der Maaten L and Weinberger K Q.2017. Densely connected convolutional networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2261-226 [DOI: 10.1109/CVPR.2017.243]

Huang H M, Lin L F, Tong R F, Hu H J, Zhang Q W, Iwamoto Y, Han X H, Chen Y W and Wu J. 2020. Unet 3+: a full-scale connected unet for medical image segmentation///Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona, Spain: IEEE: 1055-1059 [DOI: 10.1109/ICASSP40776.2020.9053405]

Jha D, Smedsrud P H, Riegler M A, Johansen D, Le Lange T, Halvorsen P and Johansen H D. 2019. ResUNet++: an advanced architecture for medical image segmentation//The 2019 IEEE International Symposium on Multimedia. San Diego, USA: IEEE: 225-2255 [DOI: 10.1109/ISM46123.2019.00049]

Ji S Y and Xiao Z Y. 2021. Integrated context and multi-scale features in thoracic organs segmentation. Journal of Image and Graphics, 26(9): 2135-2145 (吉淑滢, 肖志勇. 2021. 融合上下文和多尺度特征的胸部多器官分割. 中国图象图形学报, 26(9): 2135-2145) [DOI: 10.11834/jig.200558]

Jin S N, Zhou D B, He B and Gu J J. 2021. Segmentation of spine CT images based on multi-scale feature fusion and attention mechanism. Computer Systems and Applications, 30(10): 280-286 (金顺楠, 周迪斌, 何斌, 顾静军. 2021. 基于多尺度特征融合与注意力机制的脊柱CT图像分割. 计算机系统应用, 30(10): 280-286) [DOI: 10.15888/j.cnki.csa.008118]

Kolařík M, Burget R, Uher V, Říha K and Dutta M K. 2019. Optimized high resolution 3D dense-U-Net network for brain and spine segmentation. Applied Sciences, 9 (3): #404 [DOI: 10.3390/app9030404]

Li X and He J. 2018. Application of 3D fully convolution network in spine segmentation. Electronic Science and Technology, 31(11): 75-79 (李贤, 何洁. 2018. 3D全卷积网络在脊柱分割中的应用. 电子科技, 31(11): 75-79) [DOI: 10.16180/j.cnki.issn1007-7820.2018.11.019]

Li X M, Chen H, Qi X J, Dou Q, Fu C W and Heng P A. 2018. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Transactions on Medical Imaging, 37(12): 2663-2674 [DOI: 10.1109/TMI.2018.2845918]

Liu Y, Wu R R, Tang L and Song N N. 2022. U-Net-based mediastinal lymph node segmentation method in bronchial ultrasound elastic images. Journal of Image and Graphics, 27(10): 3082-3091 (刘羽, 吴蓉蓉, 唐璐, 宋宁宁. 2022. U-Net支气管超声弹性图像纵膈淋巴结分割. 中国图象图形学报, 27(10): 3082-3091) [DOI: 10.11834/jig.210225]

Liu Z L, Chen G, Shan Z Y and Jang X Q. 2018. Segmentation of spine CT image based on deep learning. Computer Applications and Software, 35(10): 200-204, 273 (刘忠利, 陈光, 单志勇, 蒋学芹. 2018. 基于深度学习的脊柱CT图像分割. 计算机应用与软件, 35(10): 200-204), 273) [DOI: 10.3969/j.issn.1000-386x.2018.10.036]

Löffler M T, Sekuboyina A, Jacob A, Grau A L, Scharr A, El Husseini M, Kallweit M, Zimmer C, Baum T and Kirschke J S. 2020. A vertebral segmentation dataset with fracture grading. Radiology: Artificial Intelligence, 2(4): #190138 [DOI: 10.1148/ryai.2020190138]

Long J, Shelhamer E and Darrell T. 2015. Fully convolutional networks for semantic segmentation//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3431-3440 [DOI: 10.1109/CVPR.2015.7298965]

Milletari F, Navab N and Ahmadi S A. 2016. V-Net: fully convolutional neural networks for volumetric medical image segmentation//Proceedings of the 14th International Conference on 3D Vision. Stanford, USA: IEEE: 565-571 [DOI: 10.1109/3DV.2016.79]

Oktay O, Schlemper J, Le Folgoc L, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla N Y, Kainz B, Glocker B and Rueckert D. 2018. Attention U-Net: learning where to look for the pancreas [EB/OL]. [2022-11-23]. https://arxiv.org/pdf/1804.03999.pdf

Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]

Sekuboyina A, Kukačka J, Kirschke J S, Menze B H and Valentinitsch A. 2018. Attention-driven deep learning for pathological spine segmentation//Proceedings of the 5th International Workshop on Computational Methods and Clinical Applications in Musculoskeletal Imaging. Quebec City, Canada: Springer: 108-119 [DOI: 10.1007/978-3-319-74113-0_10]

Shi Y Y, Wang Y Q, Wu C Y, Yeh C F, Chan J, Zhang F, Le D and Seltzer M. 2021. Emformer: efficient memory Transformer based acoustic model for low latency streaming speech recognition//Proceedings of the ICASSP 2021-2021 IEEE International Conference

on Acoustics, Speech and Signal Processing. Toronto, Canada: IEEE: 6783-6787 [DOI: 10.1109/ICASSP39728.2021.9414560]

Strudel R, Garcia R, Laptev I and Schmid C. 2021. Segmenter: Transformer for semantic segmentation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 7242-7252 [DOI: 10.1109/ICCV48922.2021.00717]

Tian F Y, Zhou M Q, Yan F, Fan L and Geng G H. 2020. Spinal CT segmentation based on AttentionNet and DenseUnet. Laser and Optoelectronics Progress, 57(20): #201008 (田丰源, 周明全, 闫峰, 范力, 耿国华. 2020. 基于AttentionNet和DenseUnet的脊椎CT分割. 激光与光电子学进展, 57(20): #201008) [DOI: 10.3788/LOP57.201008]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010.

Wang W X, Chen C, Ding M, Yu H, Zha S and Li J Y. 2021. TransBTS: multimodal brain tumor segmentation using Transformer//Proceedings of the 24th International Conference on Medical Image Computing and Computer-Assisted Intervention. Strasbourg, France: Springer: 109-119 [DOI: 10.1007/978-3-030-87193-2_11]

Xiao X, Lian S, Luo Z M and Li S Z. 2018. Weighted res-UNet for high-quality retina vessel segmentation//Proceedings of the 9th International Conference on Information Technology in Medicine and Education. Hangzhou, China: IEEE: 327-331 [DOI: 10.1109/ITME.2018.00080]

Zhao Y M, Li Q and Guan X. 2020. Lightweight brain tumor segmentation algorithm based on a group convolutional neural network. Journal of Image and Graphics, 25(10): 2159-2170 (赵奕名, 李锵, 关欣. 2020. 组卷积轻量级脑肿瘤分割网络. 中国图象图形学报, 25(10): 2159-2170) [DOI: 10.11834/jig.200247]

Zhou H Y, Guo J S, Zhang Y H, Yu L Q, Wang L S and Yu Y Z. 2022. nnFormer: interleaved Transformer for volumetric segmentation [EB/OL]. [2022-11-23]. https://arxiv.org/pdf/2109.03201v1.pdf

Zhou Z W, Siddiquee M M R, Tajbakhsh N and Liang J M. 2018. Unet++: a nested u-net architecture for medical image segmentation//Proceedings of the 4th International Workshop on Deep Learning in Medical Image Analysis. Granada, Spain: Springer: 3-11 [DOI: 10.1007/978-3-030-00889-5_1]

作者简介

卢玲,女,硕士研究生,主要研究方向为医学图像分割和计算机视觉。E-mail: LilyOne@stu.jhun.edu.cn

漆为民,通信作者,男,教授,硕士生导师,主要研究方向为嵌入式系统、智能控制与应用。E-mail: qwmin@jhun.edu.cn