计算机科学
COMPUTER SCIENCE

一种融合CNN和Swin Transformer的医学显微图像分割模型

孙开鑫，刘斌，苏曙光

引用本文

孙开鑫，刘斌，苏曙光. 一种融合CNN和Swin Transformer的医学显微图像分割模型[J]. 计算机科学, 2023, 50(11A): 230200119-8.

SUN Kaixin, LIU Bin, SU Shuguang. Medical Microscopic Image Segmentation Model Based on CNN Structure and Swin Transformer [J]. Computer Science, 2023, 50(11A): 230200119-8.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230200119

一种融合 CNN 和 Swin Transformer 的医学显微图像分割模型

孙开鑫　刘　斌　苏曙光
华中科技大学软件学院　武汉 430074
(sunkaixin_1009@163.com)

摘　要　医学显微图像分割在临床诊断和病理分析中具有重要应用价值。然而,由于显微图像具有形状、纹理、大小等复杂的视觉特征,因此要精确分割显微图像是一项困难的任务。文中提出了一种新的分割模型 UMSTC,该模型基于 U 型结构,并通过将 U-net 模型和 Swin Transformer 模型进行融合来兼顾图像的细节特征和宏观特征,并保持建模完整性。具体来说,UMSTC 模型的下采样部分采用 Swin Transformer 网络来优化其内含的注意力机制,以提取微观和宏观特征;上采样部分基于 CNN 网络反卷积操作,并通过残差机制接收和融合下采样阶段的特征图,以减小图像合成精度损失。实验结果表明,所提出的 UMSTC 分割模型比目前主流的医学图像语义分割模型具有更好的分割效果,其中 mPA 提高了约 3%～5%,mIoU 提高了约 3%～8%,且分割结果具有更高的主观视觉质量和更少的噪点。因此,UMSTC 模型在医学显微图像分割领域具有广泛的应用前景。

关键词:显微图像分割;Swin Transformer;卷积神经网络;注意力机制;残差网络
中图法分类号　TP391.1

Medical Microscopic Image Segmentation Model Based on CNN Structure and Swin Transformer

SUN Kaixin,LIU Bin and SU Shuguang
School of Software Engineer,Huazhong University of Science and Technology,Wuhan 430074,China

Abstract　Medical microscopic image segmentation has important application value in clinical diagnosis and pathological analysis. However,due to the complex visual features such as shape,texture,and size of microscopic images,accurate segmentation of these images is a challenging task. In this paper,we propose a new segmentation model called UMSTC,which is based on a U-shaped structure and combines the U-Net model and Swin Transformer model to balance the details and macro features of images while maintaining modeling integrity. Specifically,the down-sampling part of the UMSTC model uses the Swin Transformer network to optimize its inherent attention mechanism for extracting micro and macro features,while the up-sampling part is based on a CNN network's deconvolution operation and uses a residual mechanism to receive and fuse feature maps from the down-sampling stage to reduce image synthesis accuracy loss. Experimental results show that the proposed UMSTC segmentation model has better segmentation performance than current mainstream medical image semantic segmentation models,with mPA and mIoU increases by approximately 3%～5% and 3%～8%,respectively,and the segmentation results have higher subjective visual quality and fewer artifacts. Therefore,the UMSTC model has broad application prospects in the field of medical microscopic image segmentation.

Keywords　Microscopic image segmentation,Swin Transformer,CNN,Attention mechanism,Residual network

1　引言

显微图像分析在临床诊疗和病理分析中的应用都十分广泛。本文研究的显微图像来自法医做“是否存在弥漫性轴索损伤”病理分析的样本数据库。弥漫性轴索损伤是脑部受到严重外伤致死时在脑干切片的显微图像上可观察到的一种典型损伤。是否存在弥漫性轴索损伤是法医判定死者头部是否遭受过外力击打的重要依据。但是由于切片方式、切片角度、染色方式等不同,同一病患组织的显微图像呈现出来的视觉特征有很大的差别。因此,显微图像的分析比 X 光或 MRI 核磁共振等宏观范畴的图像分析要复杂很多[1-2]。

因此,利用 AI 辅助医学专家分析显微图像具有很强的理论价值和应用价值[3-5]。显微图像中目标的视觉特征表现十分复杂,即便是同一张显微图像中同类目标的形状、纹理、颜色、大小也有显著区别。实际中这也经常导致不同的医学专家对同一张图像的判定出现不同的结论,甚至是相反的结论。实验表明,显微图像的分析既要关注图像的细节特征,还要同时关注较大区域的宏观特征,考虑同一目标被分割到不同图像块导致目标建模不完整等问题。针对以上问题,本文方法主要有以下优势:

1)本文提出的 UMSTC 模型利用 U-net 和 Swin Transformer 模型的优势进行了融合,既兼顾了细节特征,也考虑了宏观特征,从而实现了更好的显微图像分割效果。

2)UMSTC 模型的下采样部分利用 Swin Transformer

基金项目:武汉市科技计划基金(2019010701011385)
This work was supported by the Wuhan Science and Technology Research Program(2019010701011385).
通信作者:苏曙光(sueagle@163.com)

网络的注意力机制，相对于卷积神经网络(CNN)能够更好地捕捉图像的宏观视觉特征，也更有利于分割任务的多模态学习。

3)通过实验结果表明，UMSTC 模型相对于其他医学图像分割模型如 TransUnet 在多项评估指标上表现更优，且拥有更高的主观视觉质量和较少的噪点。

本节第 2 章介绍了 CNN 和 Transformer 在图像分割领域的研究进展和相关研究基础；第 3 章介绍了新模型的结构和关键原理，包括下采样阶段、上采样阶段以及两个阶段之间的残差连接机制；第 4 章介绍了实验和结果分析；最后总结全文。

2 相关研究

卷积神经网络 CNN 在医学图像分割中应用十分广泛[6-8]。在实践中，医学图像分割任务往往面临数据集的数量较少、类别失衡和标注困难等难题。因此在医学图像分割任务中对网络模型的拟合能力和鲁棒性都提出了更高的要求[9-11]。面对显微图像这一更具挑战性的医学图像类别，目前针对医学显微图像的语义分割任务中，主要包括基于 CNN 结构的处理模型和基于 Transformer 结构的处理模型。

基于 CNN 结构的处理模型中具有代表性的有 2015 年 Ronneberger 等提出的 U-Net 模型[12]，其采用了编解码器的结构来处理医学图像分割问题，其中编码阶段采用 CNN 下采样层提取抽象特征，解码阶段则采用上采样层恢复特征图尺寸，整体结构呈现 U 字型。因为多级尺度特征图融合的收敛速度快，目前被大量应用于数据量不足的二分类分割任务中。正因为 U-Net 模型结构在医学图像领域的巨大成功，相关的变体结构也层出不穷，如 Zhou 等提出的 U-Net++模型结构[13]和 Diakogiannis 等提出的 ResU-Net 模型结构[14]。

基于 Transformer 结构[15-16]中具有代表性的如 Chen 等提出的使用 Transformer 模型与 CNN 模型相结合的 TransUnet 语义分割模型结构[17]，利用 Transformer 编码器与 CNN 网络融合作为分割网络提取特征的模块，此结构在一定程度上融合了 Transformer 和 CNN 各自在建模上的优势，使得模型能更好地对医学显微图像进行分割处理。基于此思路也诞生了更多的网络结构，如 Cao 等提出的 Swin-UNet 模型结构和 Zhang 等提出的 TransFuse 模型结构[18]。

3 UMSTC 模型的网络结构与原理

3.1 UMSTC 模型的网络结构

图 1 给出了 UMSTC 模型的整体结构，类似 U-Net 网络具有 U 形结构，逻辑上分为下采样阶段、上采样阶段以及残差连接等 3 个部分。下采样阶段采用了 Swin Transformer 结构，上采样阶段采样了 CNN 反卷积操作，残差连接用于连接下采样和上采样两个阶段，其中也具有 CNN 典型的卷积操作。

与传统的基于 CNN 卷积网络的分割模型 U-Net 相比，新模型 UMSTC 的结构有两点较大的区别。1)下采样阶段的网络结构不同。U-Net 分割模型的下采样阶段由传统的卷积和池化多次交叠构成，缺少对多尺度特征融合的支持，而 UMSTC 模型的下采样阶段采用了支持注意力机制、支持多尺度特征融合和支持更大感受野的 Swin Transformer 结构[19]。2)上采样和下采样之间的连接方式不同。U-Net 模型仅通过简单的连接把下采样阶段的特征图像传递给上采样阶段，而新模型 UMSTC 在下采样和上采样两个过程中引入了残差连接机制，这样可以进一步提高图像的分割精度，也能加快模型的收敛，避免网络退化。

图 1 给出了 UMSTC 模型，在逻辑上由下采样阶段、上采样阶段以及残差连接等 3 个部分组成。下采样阶段由多层结构相同的层构成，每层都包括特征图预处理模块和 SwinT 下采样模块。特征图预处理模块对特征图(也包括最初输入的图像)进行分片(patch)处理；SwinT 下采样模块利用注意力机制完成特征提取并完成特征图的下采样，该模块的基本工作原理参考了 Swin Transformer 模型；上采样阶段也是由多层结构相同 CNN 反卷积层构成，通过反卷积操作完成特征图的上采样；残差连接用于连接下采样阶段与上采样阶段相同的层，将下采样某一层产生的某分辨率的特征图连接到上采样相应的层进行融合，以减小图像合成的精度损失。上采样阶段的输出是与原图分辨率相同的特征图，该特征图用于实现像素级的图像分类，即图像像素级的分割。

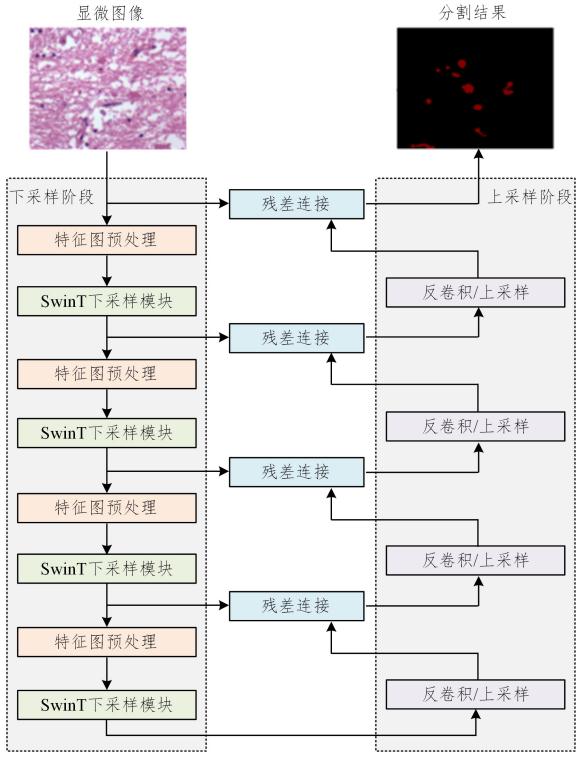


图 1 新分割模型 UMSTC 的结构

Fig. 1 Structure of new model UMSTC

3.2 特征图预处理模块

特征图预处理在下采样阶段的每一层中都会出现，其作用有两点：(1)满足 Swin Transformer 结构注意力(attention)机制的需求，把图像二维数据展平并编码为一维数据；(2)重排被分割的图像块(包括特征图)，确保图像中较大目标被分割后仍具有良好的建模完整性。

若直接将图像二维数据展平至一维序列数据，则会导致序列长度过长，进而导致模型进行注意力机制计算时复杂度过高。UMSTC 模型将图像切分成多个 16×16 的图片块后进行展平。假设原始图像的尺寸为 224×224，那么可得到 196 个尺寸一致的图像块，所有图像块经过展平后得到的序列长度为 256。

直接把原始图像切割为 16×16 图片块并展平为一维序列的方法虽然能较大程度地降低注意力机制计算的复杂度，

但是在切分后无法保证显微图像中目标信息的建模完整性。因为对于医学显微图像数据而言，同一切片拍摄出来的多张显微图像相互之间在宏观上高度相似，纹理和形状差别并不大。所以要更精确地分割目标，则务必考虑目标的完整性，也就是要考虑同一目标被人为分割到相邻的不同图像块而导致的目标不完整问题。为解决目标完整性的问题，UMSTC 模型在特征图（含原图像）预处理过程中对分割后的图像块序列做了转换，以便恢复原本相邻图像块之间可能被分割的目标的完整性。

图 2 通过示例给出了图像块序列被重排，以获得更完整目标特征的流程，序列重排的流程包括 3 步。第一步，把切割后的图像块（例如 9 块 16×16 的图像块）进行简单的线性投影（MLP 层），展平为一维序列块；第二步，把展平后的一维序列簇重建（利用 Reshape 方法）为二维序列数据；第三步，对重建后的二维序列数据进行切分和重组（利用 Unfold 算子）。每个图像块对应的数据被切分为多个部分，然后按空间相邻原则将切分后的数据重组构成新的向量。注意，重组时原本空间上相邻的图像块被组合在一起，而且重组仅做拼接。重组后的一维序列进行维度颠倒后传入 UMSTC 模型的下采样阶段（也即编码阶段）进行处理。

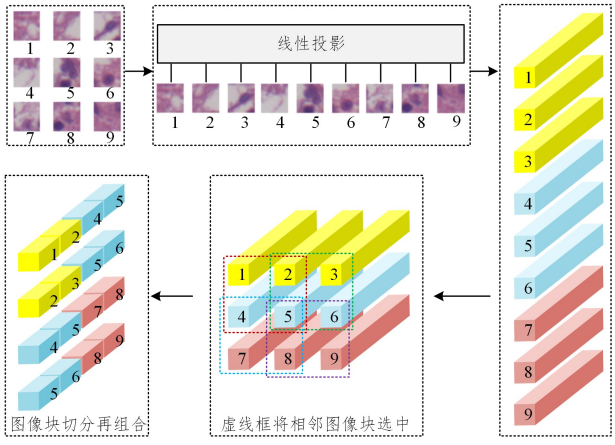


图 2　图像块序列被切分并重排

Fig. 2　Flow of image blocks sequence re-arrange

图像块序列重排不仅提高了图像中目标的完整性，也提高了模型的并行运算性能。在下采样的过程中，每一层在提取特征之前都要使用特征图进行预处理，保证图像中目标的建模完整性。这不仅降低了后续编码阶段的注意力计算复杂度，也避免了图像中同一目标因为被分割到多个相邻图像块而带来的目标建模精度损失。

3.3　SwinT 下采样模块

SwinT 下采样模块在下采样阶段的每一层中都会出现，其作用有两个：(1)通过注意力机制提取特征生成多尺度特征图；(2)执行下采样生成更低分辨率的特征图。SwinT 下采样模块的输入是本层特征图预处理模块处理后的特征图，输出则是分辨率减半的特征图，并继续送往下一层。

SwinT 下采样模块的结构如图 3 所示，采用 Swin-Transfomer 模型的编码结构，按数据流的方向先后经过 3 个子模块：窗口注意力计算模块，移动窗口注意力计算模块和下采样模块。SwinT 下采样模块的输入是上一层输出的特征图 X_i（已被本层预处理过），输出则是分辨率减半的特征图 X_{i+1}，X_{i+1} 会被继续送往下一层重复同样的过程。

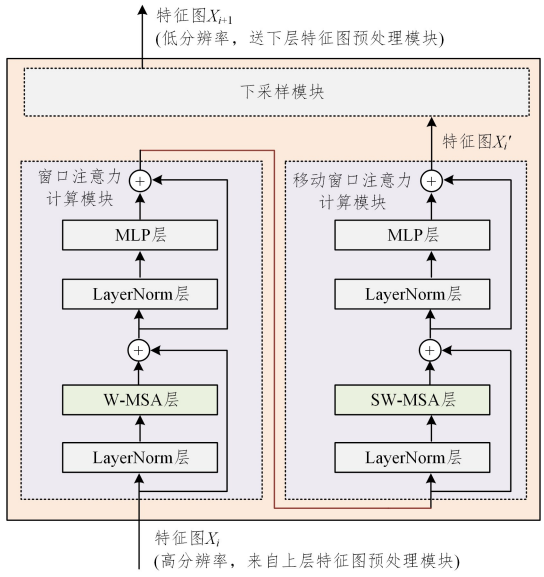


图 3　下采样阶段每层都含有的 SwinT 下采样模块结构

Fig. 3　Structure of SwinT down sampling process

3.3.1　窗口注意力计算子模块

在主流的 Transform 图像处理框架中（例如 ViT 框架），由于对图像块的划分都是固定尺寸的，因此生成的 tokens 不具有多尺度特征。为了解决这个问题，UMSTC 模型通过窗口机制将图像或特征图划分为不同尺寸的窗口，并通过窗口注意力机制对同一窗口中的所有 tokens 进行注意力计算。此外，采用窗口注意力机制相对于传统 Transform 图像处理框架在整个图像范围进行注意力计算，不仅降低了计算复杂度，也提升了计算并行度。

图 3 中左边虚线框内给出了窗口注意力计算模块的构成和其计算注意力的过程。第一步，经过前述预处理模块处理后的图像数据（或特征图数据）先经过 LayerNorm 层进行归一化处理；第二步，通过 W-MSA 层（Window Multi-head Self Attention）进行窗口注意力计算；第三步，再次通过 LayerNorm 层进行归一化；第四步，通过 MLP 层（多层感知机层）进行线性变换，处理完成该层窗口的注意力计算。

窗口注意力计算采用式(1)进行，其中 Q,K,V 为输入 tokens 序列经过 3 个不同的 MLP 层计算得到的，反映一个图像块与其他图像块之间，的联系。B 为 token 的相对位置编码数据，显然相对位置越远的图像块之间，其相关性或重要性更低，关于位置编码后文将进行描述。

$$Attention(Q,K,V)=\mathrm{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}+B\right)V \tag{1}$$

在分割显微图像时，分割出来的图像块在序列中的排放位置隐含了不同图像块的前后空间相邻关系。图像块（也即 tokens）的位置信息是每个 tokens 在序列中（即句子中或图像块序列中）的相对位置，能帮助编码器维护序列正确的语法结构，以保证语义表达的正确性。因此，UMSTC 模型也引入了位置编码（Positional Encoding），即对每个分割出来的图像块（有相应的 token 与之对应）在图像中的坐标信息进行编码。位置编码的方式有多种，如一维编码、二维编码、相对位置编码等。经过实验，UMSTC 模型采用了效果最优[19]的相对位置编码方式来记录，并在进行注意力计算时加入了位置编码。式(1)中 B 即为相对位置编码数据。在具体实现时采用相对

位置编码表和相对位置编码索引的方式实现。式(1)可实现带相对位置编码的单头注意力计算,对于多头注意力则需要使用多组单头注意力计算后进行融合。

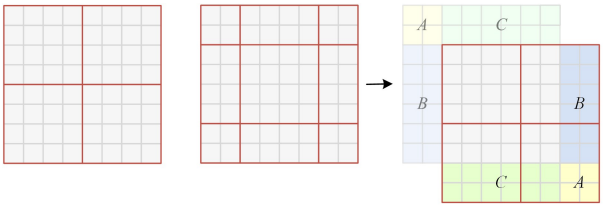
3.3.2 移动窗口注意力计算子模块

移动窗口注意力计算是 UMSTC 模型在前述窗口注意力计算环节之后的环节。采用移动窗口注意力机制可以扩大感受野,以提取更大范围内的宏观特征,从而满足显微图像的目标特点。

前述窗口注意力计算的特点是每次参与注意力计算的 tokens 集合都仅限于窗口中的图像块,而窗口之外的更多图像块被忽略了。因此,窗口注意力机制虽然提高了模型计算时的并行度和速度,但其缺陷也很明显,就是每个窗口中的图像块 tokens 序列无法与其他窗口中的图像块融合起来进行注意力运算,从而导致提取到的特征缺乏宏观性。移动窗口注意力机制通过移动窗口将多个相邻窗口关联起来,实现不同窗口之间的信息融合,从而实现全局性建模。图 3 中右边的虚线框中展示了移动窗口注意力机制的构成。移动窗口注意力计算过程与窗口注意力机制的总体流程相似,但前者用 SW-WSA 层(Shift Window Multi-head Self Attention)代替了后者的 W-WSA 层。

下文重点讨论移动窗口注意力计算过程中移动窗口的过程和窗口重新拼接的问题。在计算多头注意力时,窗口会在原有窗口划分的基础上同时向右和向下移动一定的距离(例如向下和向右各 2 个像素),但是这个操作会导致重新划分的窗口大小和形状不一致。对大小或形状不一致的窗口进行注意力计算会降低计算性能,因此需要采用一定的方法对重新划分的窗口进行重排和拼接,以获得大小一致的窗口。图 4 中,用一个 8×8 的图像(窗口大小是 4×4)作为例子来展示窗口移动前后窗口的划分情况以及对划分出来的大小和形状不一致的窗口进行重排和拼接的过程。

图 4(a)是窗口移动之前的 8×8 图像(或特征图),已被划分为 4 个完全一样的 4×4 窗口。图 4(b)是把窗口同时向右和向下各移动 2 个像素后的窗口划分情况,此时图像被划分为 9 个大小不一致的窗口。若模型直接对图 4(b)中的各个窗口进行注意力计算,将难以利用硬件平台的并行计算能力。图 4(c)是把图 4(b)中不规则的窗口进行适当移位并拼接为 4×4 大小的规则窗口的过程。将不规则的 A,B,C 3 块从左上侧移动至右下侧并重新拼接,原本 9 个大小不一致的窗口重新变为 4 个大小均为 4×4 的窗口。



(a)分为 4 个大小一致的窗口 (b)分为多个大小不一致的窗口 (c)将部分窗口移动并拼成 4 个大小一致的窗口

图 4 移动窗口和窗口的重新拼接

Fig. 4 Shift windows and re-arrange windows

窗口进行拼接后虽然有利于采用并行计算提高运算速度,但是会导致原本不相邻(现在已相邻)的两个窗口块之间进行注意力计算时带来较大的误差。因此,还需要对拼接后的 4×4 图像块做一定的处理,以保证只有原本确实相邻的图像块之间才有较高的注意力值,而原本不相邻(现在已相邻)的图像块之间有较低的注意力值。因此,UMSTC 模型在计算注意力时通过加入掩码来降低原始不相邻区域的注意力值。

3.3.3 下采样子模块

下采样模块在下采样阶段中的作用有两点:(1)实现在编码阶段降低特征图分辨率,以输出多个尺度特征图;(2)确保能采用多尺度窗口对图像数据进行注意力计算。

下采样阶段中下采样模块的工作原理是:对输入的维度为[N,C,H,W]的特征数据在其 H,W 两个维度上分别以 0 和 1 为起点,以 2 为步长,进行下采样得到一系列维度为[N,$C*4$,$H/2$,$W/2$]的数据,也即获得了一系列不同尺度的特征图数据。新的特征图数据的 H 和 W 两个维度已经实现减半,即获得了低分辨率的特征图。图 5 用示例的方式给出了一个维度[1,8,8]的特征图被下采样为维度[4,4,4]的特征图的过程。UMSTC 模型通过下采样实现了在 Swin Transformer 结构中生成多尺度的特征图,有利于特征图像的多尺度特征融合,不仅提高了模型的拟合能力,也更利于模型快速收敛。

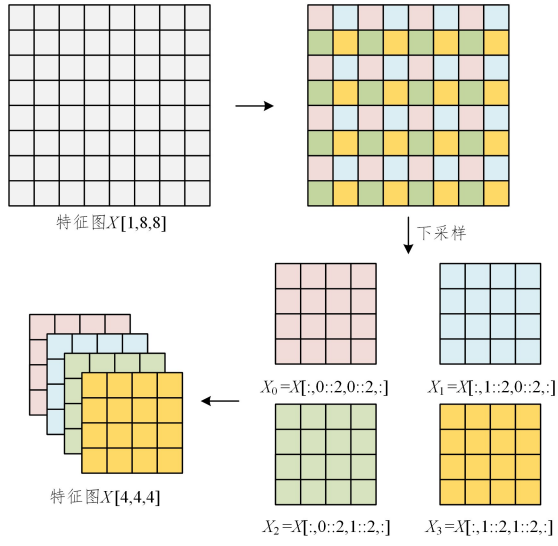


图 5 下采样的过程和示例

Fig. 5 Process and example of down sampling

需要注意的是,下采样过程中不同层产生的不同尺寸特征图将会通过残差连接机制送到上采样过程中进行后续的多尺寸特征图融合。

3.4 上采样阶段

上采样阶段,也即解码阶段,主要作用是不断提高深层低分辨率特征图的分辨率,逐层迭代以重建具有原始分辨率的特征图。图 1 给出了上采样阶段由多层构成,每层都包含一个上采样模块,通过执行反卷积的方式执行上采样。需要注意的是,UMSTC 模型还在上采样过程中,通过残差连接机制接收并融合了前述下采样阶段各层输出的相应分辨率特征图。3.4.1 节介绍上采样模块,3.4.2 节介绍残差连接机制以及其中的残差模块。

3.4.1 上采样模块

上采样模块完成深层特征图分辨率的提高,通过多层迭代,当最后一层的上采样模块输出了与原始图像尺寸一致的特征图时即可实现对图像中每个像素点的分类,也即实现了

对原图像的语义分割。上采样的方法主要有图像插值和反卷积两种。由于图像插值法没有可学习参数，UMSTC 模型采用了反卷积方法进行上采样。反卷积使用反卷积核在特征图上逐点处理，但与卷积运算不同的是反卷积运算提升了特征图的分辨率。另外，为提升运算效率，反卷积操作采用矩阵运算来完成，低分辨率特征图被展成一维，反卷积核被展成稀疏矩阵。

图 6 给出了 UMSTC 模型通过三步把一个 2×2 特征图上采样为 4×4 特征图的过程。

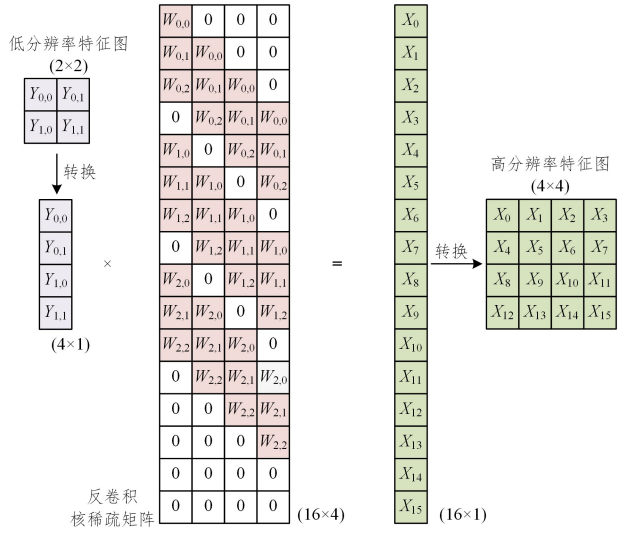


图 6　通过反卷积进行上采样的过程和示例

Fig. 6　Process and example of up sampling by deconvolution

第一步，将上一层输出的尺寸为 2×2 的特征图转化为 4×1 的特征图；第二步，将 16×4 的反卷积核稀疏矩阵与 4×1 的特征图进行矩阵乘得到 16×1 的特征图；第三步，将 16×1 的特征图转换为 4×4 的特征图。由此便完成了由低分辨率特征图到高分辨率特征图的转换。

3.4.2　利用残差连接实现特征融合

在上采样过程中为实现多尺度特征图的融合，并保证下采样阶段的特征图与上采样恢复的特征图之间的抽象性差异不大，UMSTC 模型设计了残差连接机制把下采样和上采样相同的层进行了连接。

图 7 给出了残差连接机制和其内部的残差模块（ResBlock）的构造和工作原理。残差连接机制内部包括两个残差模块。残差连接机制有两个输入和一个输出。一个输入是来自下采样过程某层（例如第 i 层）输出的特征图 X_i，另一个输入是上采样过程对应的卷积层（假如也是第 i 层）输出的特征图 F_i，输出是特征图 F_i'，3 个特征图 X_i，F_i 和 F_i 的′分辨率相同（图中忽略了通道数）。

残差连接机制把来自下采样的特征图 X_i 经过第一个残差模块处理后与来自上采样过程的特征图 F_i 进行融合（拼接即可），融合后的特征图再次通过第二个残差模块处理后输出最终的特征图 F_i'。由于残差连接机只连接下采样和上采样之间的相同层，因此特征图 X_i，F_i 和 F_i' 的长、宽两个维度都是相同的。

残差连接机制中两个相同残差模块分别在特征图 X_i 和 F_i 融合前后进行残差特征提取，这种结构不仅能让浅层特征与深层特征能有效结合，且能有效防止梯度消失等问题。

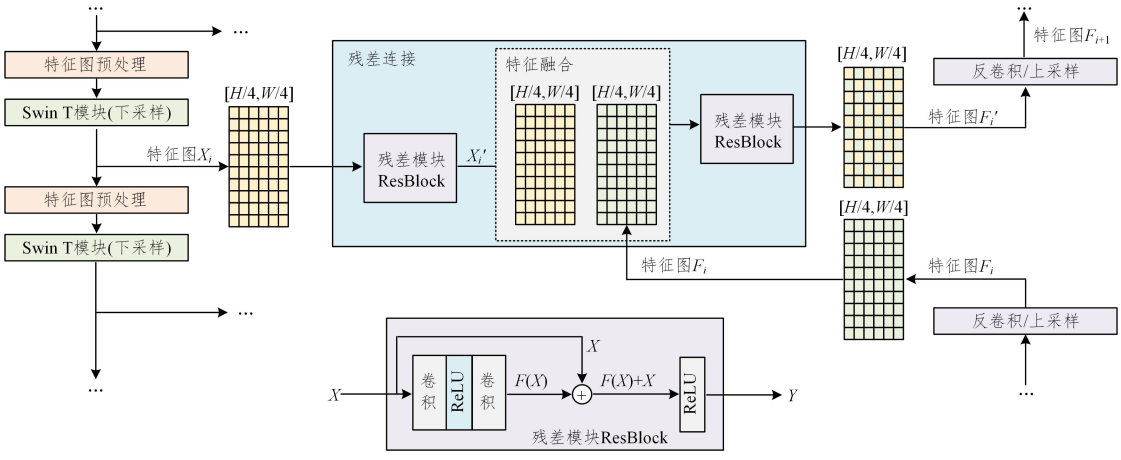


图 7　残差连接（内含残差模块）机制

Fig. 7　Mechanism of residual connection with residual module

4　实验与数据分析

4.1　实验环境

表 1 列出了实验环境的硬件、软件和开发环境的配置。

表 1　实验环境配置

Table 1　Experimentenvironment

操作系统	Ubuntu 18.04
CPU	Intel(R)Core(TM)i7-8700K CPU @3.70GHz
GPU	GeFore RTX 3090 24GB
内存	2×16GB
深度学习框架	Python 3.7+，Pytorch1.11.0，CUDA 11.1
其他依赖库	Numpy，Opencv，Logging

实验使用的操作系统为 Ubuntu 18.04 操作系统，CPU 型号为 Intel i7-8700K 3.70 GHz，GPU 型号为 GeFore RTX 3090，对应 CUDA 版本为 11.1。深度学习框架使用 Pytorch 框架，Pytorch 版本为 1.11.0，Python 版本为 3.7。

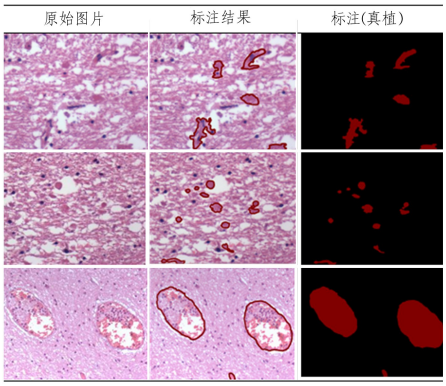
4.2　数据集与预处理

本文构建了脑部弥漫性轴索损伤显微图像分割数据集，数据集由 20 张分辨率为 50 000×50 000 的显微图像裁剪而成，并经过筛选后得到 5 200 张分辨率 224×224 的样本。样本通过 LabelMe 软件完成标注和预处理。标注时图像分割为病变区域和正常区域。表 2 列出了部分标注样本，其中第一列为原始图片，第二列为标注图片，第三列为标注掩码

(真值数据)。在模型训练时第一列作为输入图像,第三列作为真值数据,用于与模型预测输出计算交叉熵损失。基于此数据集,对 UMSTC 模型和主流分割模型进行了一系列对比实验。

表 2 实验数据集样例

Table 2 Samples of test image set



4.3 对比实验和结果分析

模型训练的样本集划分和主要参数设置为:有标注的样本集总计 5200 张,其中 5000 张以 8∶2 的比例划分为训练集和验证集,余下 200 张作为测试集,训练轮次 300 轮(epoch)。采用 SGD 优化器进行权重学习,使用正态分布初始化模型参数。初始学习率设为 0.001,第 100 轮后学习率改为 0.0001,此后每 50 轮学习率就降低为上次的 1/10。另外 momentum 设为 0.9,weight_decay 设为 10^{-8}。

图 8 给出了训练过程中损失的变化曲线。实验结果表明,训练损失和验证损失的变换趋势相同且差值绝对值较小,这表明模型训练过程中具有较好的拟合性,没有出现过拟合的情况。

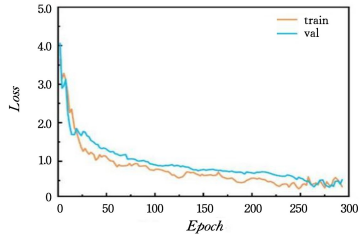


图 8 UMSTC 模型训练损失变化曲线

Fig. 8 Loss curve of training of UMSTC model

表 3 列出了 UMSTC 模型与其他主流分割模型(包括 FCN,U-Net,Resnet50,Resnet101,TransUnet 等)对脑部弥漫性轴索损伤显微图集的分割结果对比。“原始图像”列为显微图像原图,“真值 label”列为显微图像的真值标注 label 数据,其余各列是不同分割模型的分割效果,“UMSTC”列为本文设计的 UMSTC 模型的分割效果。实验结果表明,UMSTC 模型的分割效果在主观上优于其他模型,分割的边缘与真值更接近,噪点更少。

表 3 UMSTC 与主流分割模型的主观分割效果

Table 3 Experiment results of UMSTC and main segmentation models

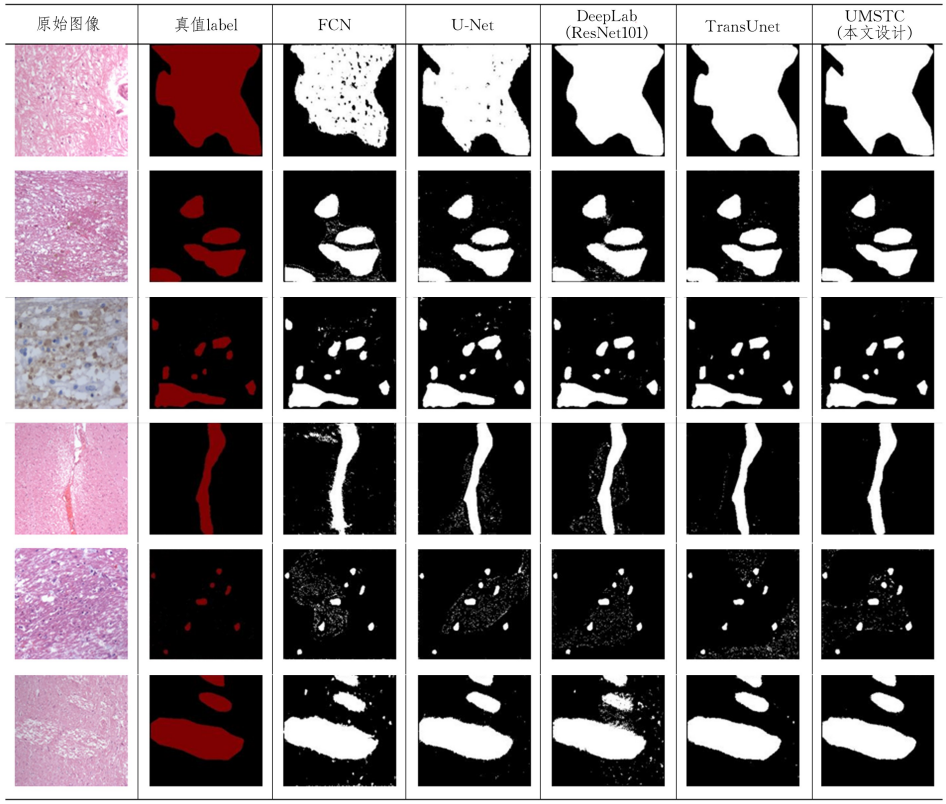


表 4 列出了不同模型分割结果的客观指标对比。对比的评估指标包括:平均像素精度(mPA)、平均交并比(mIoU)、mDice 和模型的参数量。所有模型在训练时都采用交叉熵损失函数计算模型输出值与真值(即标签 Label 值)之间的损失。实验数据表明,本文提出的 UMSTC 模型针对脑部弥漫性轴索损失类显微图像的分割效果高于其他常用分割模型。例如,在 mPA 指标上 UMSTC 模型比第 2 名 TransUnet 分割模型高约 2%,比经典的 U-Net 分割模型则高 5%。在 mIoU 指标上,UMSTC 模型比第 2 名 TransUnet 分割模型高约 3%,比经典的 U-Net 分割模型则高 8%。

表 4　UMSTC 模型与主流分割模型实验的主要客观指标对比

Table 4　Objective experiment results of UMSTC and main segmentation models

模型名	mPA/%	mIoU/%	mDice/%	Img-size	Params	FLOPs
FCN	75.98	66.45	74.90	224×224	32.9×10^6	26.6×10^9
U-Net	80.23	68.78	79.18	224×224	31×10^6	41.9×10^9
nnU-net	84.62	75.75	81.50	224×224	32×10^6	42.3×10^9
DeepLab(Resnet50)	79.45	70.90	78.34	224×224	39.6×10^6	31.4×10^9
DeepLab(Resnet101)	81.67	78.10	80.10	224×224	58.6×10^6	46.3×10^9
TransUnet	83.23	73.40	81.60	224×224	61.2×10^6	25.6×10^9
TransFuse	84.10	74.30	81.30	224×224	72×10^6	31.2×10^9
UMSTC(本文设计)	85.12	76.50	82.50	224×224	59.8×10^6	12.4×10^9

尽管 UMSTC 模型的主观分割效果和客观分割效果都较其他主流分割模型好，但是表 3 也表明，UMSTC 模型的参数量 Params 较大，这意味着 UMSTC 模型在推理过程中需要消耗更多的内存，也需要花费更多的访存时间，从而降低 UMSTC 模型的推理速度。因此，在实际应用中有必要对 UMSTC 模型进行轻量化处理，降低其参数量。

4.4　消融实验和结果分析

为了进一步论证模型结构中不同模块在整体模型中对模型分割效果的影响，本节通过多组消融实验来进行对比论证。

对模型下采样阶段中的特征图预处理阶段进行消融实验，结果如表 5 所列，为特征预处理模块消融实验，可以发现本文采用的 Tokens To Tokens 特征图预处理方法在模型精度上的表现更优。

表 5　特征图处理方法对比消融实验

Table 5　Comparison and ablation experiment of feature map processing methods

（单位：%）

特征图预处理方法	mPA	mIoU	mDice
Tokens To Tokens	85.12	76.5	82.5
Reshape	84.89	75.3	81.0
不使用特征图预处理	83.6	74.2	80.9

针对模型在编解码之间的连接方式，本节也做了进一步的消融实验，通过对比多种连接方式，以及是否采用多尺度特征图融合的方式连接等，验证不同的连接方式对模型精度对影响。

表 6　编解码连接方法对比消融实验

Table 6　Comparison and ablation experiment of encoding and decoding connection methods

（单位：%）

编解码连接方法	mPA	mIoU	mDice
多尺度残差块连接(Res Block)	85.12	76.5	82.5
残差块连接(Res Block)	82.59	73.6	78.0
多尺度直通连接	83.80	73.9	80.6
直通连接	80.99	69.8	74.2
多尺度注意力机制连接(Attention)	84.70	75.9	81.6
注意力机制连接(Attention)	81.90	74.1	78.9

通过实验对比可以发现，采用多尺度特征图的融合相对于不采用效果更优，采用残差块连接比采用直通连接和注意力机制连接的效果更优。

结束语　医学显微图像分析在临床诊疗和病理分析中的应用都十分广泛。本文以法医行业“是否存在弥漫性轴索损伤”显微图像分析为背景，研究显微图像的分割问题。显微图像中目标的视觉特征表现得十分复杂，即便是同一张显微图像中同类目标的形状、纹理、颜色、大小也有显著区别。显微图像的分析既要关注图像的细节特征，还要同时关注较大区域的宏观特征。本文结合 CNN 模型和 Swin Transformer 模型，设计了一种 U 型结构的显微图像分割模型，既能同时关注图像的细节特征和宏观特征，还能保证在同一目标被分割到不同图像块时具有较高的建模完整性，保证对显微图像实现较高精度的分割。新模型下采样阶段（即编码阶段）采用了 Swin Transformer 结构，上采样阶段（即解码阶段）采用了 CNN 反卷积操作，并且两个阶段通过残差网络进行连接，以提高分割精度。实验结果表明，新建的分割模型无论是在客观质量上，还是在主观质量都比主流的 FCN，U-Net，Resnet50，Resnet101，TransUnet 等分割模型要好，其中比 U-Net 和 TransUnet 网络的 mPA 提高了约 3%～5%，mIoU 提高约 3%～8%，且分割的主观视觉质量更高，边缘更准确清晰，噪点也更少。

参 考 文 献

[1] WANG F L. Experimental Study on Detecting Diffuse Axonal Injury with FTIR Mapping[D]. WuHan: Huazhong University of Science & Technology, 2018.

[2] LI S X. Study of diffusion tensor imaging and immunohistochemistry on diffuse axonal injury[D]. WuHan: Huazhong University of Science & Technology, 2012.

[3] LEI T, ZHOU W, ZHANG Y, et al. Lightweight v-net for liver segmentation[C]// 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020). IEEE, 2020: 1379-1383.

[4] VALANARASU J M J, PATEL V M. UNeXt: MLP-based Rapid Medical Image Segmentation Network [J]. arXiv: 2203.04967, 2022.

[5] CHEN P H C, GADEPALLI K, MACDONALD R, et al. Microscope 2.0: an augmented reality microscope with real-time artificial intelligence integration[J]. arXiv: 1812.00825, 2018.

[6] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[7] 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.

[8] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv: 1704.04861, 2017.

[9] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the

IEEE conference on computer vision and pattern recognition. 2015:3431-3440.

[10] 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.

[11] 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.

[12] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[C]// Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015. 2015: 234-241.

[13] ZHOU Z, SIDDIQUEE M M R, TAJBAKHSH N, et al. A Nested U-Net Architecture for Medical Image Segmentation[J]. arXiv: 1807. 10165, 2018.

[14] DIAKOGIANNIS F I, WALDNER F, CACCETTA P, et al. ResU-Net-a: A deep learning framework for semantic segmentation of remotely sensed data[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 162(1): 94-114.

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

[16] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv: 2010. 11929, 2020.

[17] CHEN J, LU Y, YU Q, et al. TransUnet: Transformers make strong encoders for medical image segmentation[J]. arXiv: 2102. 04306, 2021.

[18] ZHANG Y, LIU H, HU Q. Transfuse: Fusing transformers and cnns for medical image segmentation[C]// Medical Image Computing and Computer Assisted Intervention-MICCAI 2021. 2021: 14-24.

[19] 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.

SUN Kaixin, born in 1998, postgraduate. His main research interests include deep learning and image processing.

SU Shuguang, born in 1975, Ph.D, assistant professor, Ph. D supervisor, is a member of China Computer Federation. His main research interests include machine learning, image processing and embedded system.