计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(7):94−102 [doi: 10.15888/j.cnki.csa.009540]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 BM-TransUNet 的咽后壁识别分割①

王世刚, 孙静雯

(广西科技大学 自动化学院, 柳州 545616)
通信作者: 王世刚, E-mail: gxwsg@gxust.edu.cn

摘　要: 图像分割经历了从基于传统的阈值分割等方法逐步发展到基于卷积神经网络的方法. 传统的卷积神经网络在分割领域中表现突出, 但训练速度慢、分割精度不够高等局限性也逐渐显现. 为了克服这些局限性, 本文在 TransUNet 网络的基础上进行改进, 提出了基于 BM-TransUNet 网络的图像分割识别方法, 在 TransUNet 网络的在第 1 层之后加上深度可分离卷积模块, 并在编码器下采样的卷积层后引入注意力机制模块, 让算法更好地探索分割对象特征, 同时在编码器与解码器之间引入多尺度特征融合模块 FPN. 本文基于自制的咽后壁数据集, 用于图像分割训练, 并将训练后的 BM-TransUNet 网络与多种传统分割网络的效果进行对比. 实验结果表明, 相比于其他传统的深度学习模型, BM-TransUNet 网络的识别方法具有较高的分类准确性和泛化能力, 精确度 *Precision* 和 *Dice* 系数分别达到了 93.61% 和 90.76%, 显示出较好的计算效率, 能有效地应用于分割任务.

关键词: BM-TransUNet 网络; 图像分割; 注意力机制模块; 多尺度特征融合模块; 咽后壁数据集

引用格式: 王世刚,孙静雯.基于 BM-TransUNet 的咽后壁识别分割.计算机系统应用,2024,33(7):94–102. http://www.c-s-a.org.cn/1003-3254/9540.html

Posterior Pharyngeal Wall Recognition and Segmentation Based on BM-TransUNet

WANG Shi-Gang, SUN Jing-Wen

(School of Automation, Guangxi University of Science and Technology, Liuzhou 545616, China)

Abstract: Image segmentation has gradually developed from traditional threshold-based methods to convolutional neural network (CNN)-based methods. Traditional CNNs are outstanding in the field of segmentation, but the limitations of slow training speed and low segmentation accuracy are gradually emerging. To overcome these limitations, this study proposes an image segmentation recognition method based on the BM-TransUNet network, which is an improvement. A depth-separable convolution module is added to the first layer of the TransUNet network, and an attention mechanism module is introduced to the convolution layer of the encoder under-sampling so that the algorithm can better explore the features of the segmented objects. At the same time, a multi-scale feature fusion module, the feature pyramid network (FPN), is introduced between the decoder and encoder. In this study, a self-made posterior pharyngeal wall dataset is used for image segmentation training, and the effects of the trained BM-TransUNet network are compared with various traditional segmentation networks. Experimental results show that, compared to other traditional deep learning models, the identification method of the BM-TransUNet network exhibits higher classification accuracy and generalization ability, with *Precision* and *Dice* coefficient of 93.61% and 90.76%, respectively, showing better computational efficiency and effective in segmentation tasks.

Key words: BM-TransUNet network; image segmentation; attention mechanism module; multi-scale feature fusion module; posterior pharyngeal wall dataset

① 基金项目: 广西科技基地和人才专项 (AD22080004)
收稿时间: 2023-12-09; 修改时间: 2024-01-09; 采用时间: 2024-02-23; csa 在线出版时间: 2024-05-31
CNKI 网络首发时间: 2024-06-04

1 引言

深度学习 (deep learning) 是目前研究的热点领域, 与传统的人工神经网络相比, 优点是可以通过多层次的抽象从数据中自动学习特征[1]. 而基于卷积神经网络 (convolutional neural network, CNN) 的图像目标区域分割方法的快速发展, 被广泛应用到图像分割领域. 医学图像分割在计算机视觉领域具有重要的研究意义和应用价值, 它旨在利用计算机算法和技术, 对医学影像 (如 CT、MRI、X 射线等) 进行像素级别的分类和分割, 从而准确地识别和提取出感兴趣的解剖结构或病变区域. 在过去的几十年里, 图像分割领域取得了巨大的进展, 图像分割技术在医学图像分析[2]、自动驾驶[3]等领域中得到广泛应用, 但仍然存在许多挑战, 如复杂背景下的目标分割、遮挡物体的分割以及光照变化的影响等. 传统的图像分割方法通常依赖于边缘检测、区域生长或基于图论的技术, 然而这些方法往往对复杂场景表现不佳, 且对噪声敏感.

深度学习方法在图像分割任务上取得了显著的性能提升, 并成为当前主流的研究方向之一. 语义分割领域的开山之作是 Shelhamer 等人在 2015 年提出的端到端全卷积神经网络 (fully convolutional network, FCN)[4]. 与传统的卷积神经网络相比, FCN 将传统 CNN 后面的全连接层换成了卷积层, 同时为解决卷积和池化导致图像尺寸的变小, 使用上采样方式对图像尺寸进行恢复, 确保准确性和抗干扰性. 许多研究人员提出对 FCN 的改进算法, 其中最著名的是 Ronneberger 等人在 2015 年提出的 U-Net 网络[5], 它可以实现图片像素的定位, 对图像中的每一个像素点进行分类, 大大提升了医学图像分割的性能. Park 等人应用 U-Net 从高分辨率计算机断层扫描图像中分割肺部并获得较好结果[6]. 金鹭等人结合 U-Net 网络对算法进行优化改进, 用于视网膜血管分割研究, 为专家在判断病变情况时提供更多有效信息[7]. 马豪等人提出基于模型压缩与重构 U-Net 的端到端框架, 以完成实时胰腺图像分割任务[8]. Transformer 模型是由谷歌于 2017 年提出的应用于自然语言处理领域的模型框架[9], 随着研究的推进, Transformer 结构被逐步引入到计算机视觉任务中, 并与 U 型网络结合, 例如 Chen 等人提出 TransUNet (Transformers and U-Net)[10]应用于医疗图像分割; 韩文龙采用混合编码的 TransUNet 作为基线算法来提取研究区域场景中的水体信息, 来提高水体提取语义分割算法的性能[11]; Chen 等人利用 TransUNet 结合 MT 的 MT-TransUnet 多任务框架用于进行肿瘤分类以及皮肤癌识别[12]; Wang 等人提出利用 TransUNet 进行自提取变换的三维医学图像分割, 该方法可以同时学习全局语义信息和局部空间细节特征[13]; Chang 等人结合 TransUNet 对肺肿块进行分割, 有效减少不同放射科医生在诊断中造成的误差, 帮助他们提高决策的稳定性, 提高了分割准确率[14].

TransUNet 网络具备同时降低计算负担和有效捕捉重要信息的特性, 目前在分割领域较为出色, 但该算法在处理口腔内部较复杂环境且对光滑易反光的咽后壁分割时, 会出现若干问题: M 区域识别不完整, 内壁光滑导致识别分割不清楚、无法识别出完整的悬雍垂, 以及模型泛化力不足而难以映射多尺度信息等. 针对这些问题, 本文对 TransUNet 网络模型进行改进: 在输入特征图上先进行局部特征提取和压缩, 再应用注意力机制来整合全局信息和局部信息, 最后通过多尺度特征融合模块进行特征融合: 1) 在网络的第 1 层之后加上深度可分离卷积模块, 用于局部特征提取和压缩, 能够更好地平衡参数数量和模型性能; 2) 在每个下采样的卷积层都引入注意力模块 CBAM, 可以使模型在保留较大感受野的同时, 对特征进行通道和空间注意力调整, 从而提高模型对重要特征的关注度, 并进一步增强特征的建模能力; 3) 在编码器和解码器之间添加多尺度特征融合模块, 更好地捕捉不同尺度的特征, 从而提高模型的准确性和鲁棒性.

2 网络结构与改进方法

2.1 TransUNet 网络模型

在 TransUNet 网络中, 为了实现更好的性能, 采用了 R50-ViT-B_16 模型作为基础. 该模块是在 ViT (Vision Transformer)[15]的基础上引入 ResNet50[16]构成的特征提取器. ResNet50 能够从图像中提取出丰富的低级特征, 其中包含图像的细微结构和纹理信息, 提供了更强大的特征提取能力. 当输入一张图片 $x \in R^{H\times W\times C}$ 时, 先进行高级特征的提取, 通过 ViT 模型获取图像中的全局特征, 结合低级特征与 ViT 模型提取的高级特征, 并通过跨层连接, 将特征提取器获取的中低级特征图与解码模块生成的特征图进行连接, 进一步提升了网络的性能和准确性.

批嵌入模块 (patch embedding) 的主要功能是将输入的图像维度转化为一系列具有位置信息的序列. 这

种转换的好处是可以实现并行处理, 并有效利用全局信息. 对于输入图片, 网络首先对图片进行标记化, 将输入的图片$\{x \in R^{H\times W\times C}\}$重塑为一系列 2D 的 patch 切片序列$\{x_p^i \in R^{p^2C} | i=1,\cdots,N\}$. 其中, $H\times W$ 代表图片的分辨率, C 代表图片的通道数, P 代表 patches, $N=\dfrac{HW}{P^2}$ 是切片 patches 的数量, P^2 为每个 patch 的大小, 如式 (1) 所示:

$$z_o=\left[x_p^1E;x_p^2E;\cdots;x_p^NE\right]+E_{\text{pos}} \tag{1}$$

其中, O 是批嵌入模块的位置编码操作, $E\in R\left(P^2\cdot C\right)^{\times D}$ 是切片线性投影, $E_{\text{pos}}\in R^{N\times D}$ 表示位置信息的投影.

Transformer 能处理不确定长度输入的序列到序列 (Seq2Seq). 它的主要结构是编码器-解码器结构 (encoder-decoder structure), 在编码器中进行序列的编码, 并在解码器中生成相应的输出序列, 如图 1 所示.

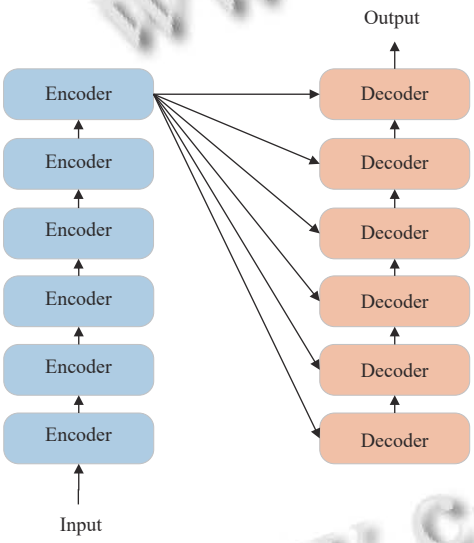


图 1 编码器-解码器结构

Transformer 模块包含了 12 个 Transformer 网络层, 每层编码器由多头自注意力机制 (multi-heads self-attention, *MSA*)、多层感知机 (multi-layer-perceptron, *MLP*) 和正则化层 (layer normalization, *LN*) 组成. 经过 patch embedding 处理后的序列经过每层 *MSA* 和 *MLP* 后的输出如式 (2)、式 (3):

$$z_l'=MSA(LN(z_{l-1}))+z_{l-1} \tag{2}$$

$$z_l=MLP(LN(z_l'))+z_l' \tag{3}$$

其中, $LN(\cdot)$ 表示层归一化算子, z_{l-1} 为 l 层的前一层 Transformer 的输出, z_l 为第 l 层的 Transformer 的输出, 并作为下一层的输入, z_l' 为 *MSA* 的输出与 z_{l-1} 的残差连接. Transformer 层的结构如图 2 所示.

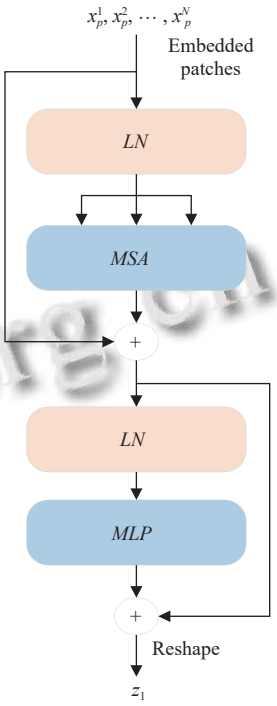


图 2 Transformer 层结构

TransUNet 采用混合 CNN-Transformer 架构作为编码器来实现精确定位[17]. 在多头自注意力机制中, 多个 q、k、v 向量分别构成矩阵 Q、K、V, 将每个组合的参数分解到不同的子空间中, 以计算注意力权重.通过多次并行计算后, 再将结果在通道维度上进行拼接, 合并得到所有子空间中的注意力信息. 其中 Q 矩阵是查询矩阵, K 矩阵是键值矩阵, V 是值矩阵. 其结构如图 3 所示.

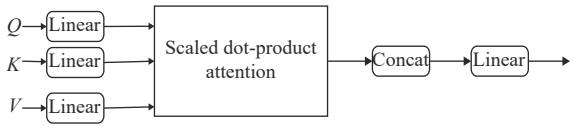


图 3 多头自注意力机制

级联上采样器 (CUP) 通常由多个上采样模块组成, 每个模块都负责将特征图进行上采样, 并与来自下一级的特征图进行融合. 这种级联结构允许逐渐增加分辨率, 并逐步融合多尺度的信息. 对 Transformer 输出的特征序列$z_L\in R^{\frac{HW}{P^2}\times D}$的形状重塑为$\dfrac{H}{P}\times\dfrac{H}{P}\times D$, 通过 4 个上采样块来将 CUP 实例化. 将重塑后的特征图依次通过 3 次上采样来获得不同维度的特征图, 并利

用跳跃连接 (skip connection) 实现了不同分辨率级别的特征融合, 恢复因下采样而丢失的中低级细节信息.

TransUNet 使得每个特征解码器用了双线性插值方法 (bilinear interpolation) , 以恢复分割结果作为上采样方式, 将低分辨率的特征图上采样到与原始输入图像相同的尺寸, 以恢复分割结果. 另外, 特征编码器的输出与特征解码器的输出通过跳跃连接相连, 以保留更多的细节信息, 达到更好的分割效果.

2.2 改进方法

2.2.1 引入深度可分离卷积模块

深度可分离卷积 (deep separable convolution, DSC) 通常作为卷积神经网络模型的一部分使用, 它可以替代传统的标准卷积层, 用于构建更轻量级、计算效率更高的神经网络结构, 允许独立学习方向和空间信息, 从而保证旋转和平移不变性, 同时, 在各种尺度上实现全局不变性, 简化了可训练参数的数量, 从而大幅减少了训练过程中的计算负担, 而且并不会降低其准确性. 具体来说, 深度可分离卷积由两部分组成: 逐深度卷积 (depthwise convolution, DW) 和逐点卷积 (pointwise convolution, PW). 逐深度卷积负责处理输入数据的各个通道, 将通道拆分, 每个通道用单独的卷积核卷积, 即输入特征的通道个数和卷积核的个数一致, 并且每个卷积核只有一个通道. 而逐点卷积则用来整合各个通道的信息, 卷积核通道数和输入的通道数相同.

本文将深度可分离卷积模块加在初始卷积层后, 也就是网络的第 1 层之后, 用于局部特征提取和压缩, 能够更好地平衡参数数量和模型性能, 有助于提高网络的表示能力, 并且能够降低计算成本.

2.2.2 增加注意力机制模块

注意力机制主要包括利用空间注意力和通道注意力, 卷积注意模块 (convolutional block attention module, CBAM) 使用串联和并行两种模块的方式, 将通道注意力图和空间注意力图分开, 引入全局池来利用空间全局信息. 通道注意力模块 (channel attention module, CAM) 是分析图像中不同通道的重要程度. 对输入特征图的每个通道执行两个并存的最大池化层和平均池化层, 再计算每个通道上的最大特征值和平均特征值, 再将特征向量输入到共享全连接层中以学习每个通道的注意力权重, 再将结果输入到激活函数, 获得了输入特征层中每个通道的权值. 获得权值后, 乘上原输入特征层, 得到注意力加权后的通道特征图. 图 4 为通道注意力模块结构图.

$$M_c(F)=\sigma(MLP(AvgPool(F))+MLP(MaxPool(F))) \tag{4}$$

其中, 特征图$F\in R^{C\times H\times W}$, σ表示激活函数 Sigmoid.

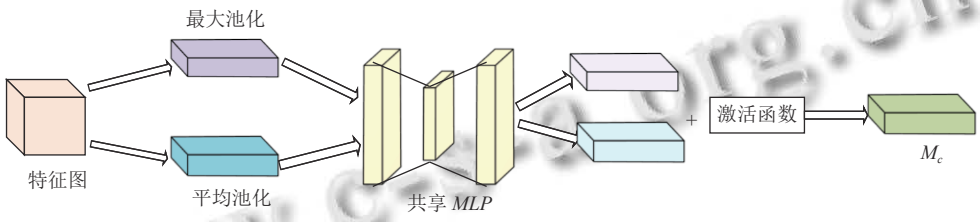


图 4 通道注意力模块

空间注意力模块 (spatial attention module, SAM) 是寻找图像中对结果影响更重要的区域, 即获取区域的权重. 将输入特征图最大池化和平均池化后的特征沿着通道维度进行连接, 接着通过卷积层处理这个特征图, 以生成空间注意力权重. 图 5 为空间注意力模块结构图.

$$M_s(F')=\sigma(f([AvgPool(F');MaxPool(F')])) \tag{5}$$

其中, 特征图$F'\in R^{C\times H\times W}$, σ表示激活函数 Sigmoid.

为使 TransUNet 算法更快速地检测到特征, 在每个下采样的卷积层后都引入一个轻量级的 CBAM 注意力模块, 产生的计算几乎可以不计入考虑, 其结果可以使算法特征更加敏感, 提高提取性能. 以使模型在保留较大感受野的同时, 对特征进行通道和空间注意力调整, 同时能够自适应地学习不同通道和空间位置的重要性, 从而提高模型对重要特征的关注度, 并进一步增强特征的建模能力.

2.2.3 引入多尺度特征融合结构

在咽后壁分割任务中, 由于咽后壁的形态和结构多样性较大, M 区域的大小存在不确定性, 单一尺度的特征可能无法完全捕捉到目标的细节和上下文信息.

多尺度特征融合模块 FPN (feature pyramid network) 可以同时处理来自不同尺度的特征并将它们进行融合, 这样可以使模型能够从不同层次、不同尺度的特征中获取关键的上下文信息, 有助于改善分割咽后壁 M 区域的准确性和鲁棒性. FPN 采用自顶向下的路径和横向连接的方式进行特征融合, 其中: 在自顶向下过程中, FPN 模块使用上采样操作来将低分辨率的特征图上采样到高分辨率, 同时进行特征融合, 得到更加丰富的多尺度特征图; 而在横向连接中, FPN 模块将高分辨率的特征图与低分辨率的特征图进行连接, 以利用低层特征的细节信息. 在合并全部分支的特征图后, 使用卷积层进行特征融合和细化, 将融合后的特征图输入到解码器中进行进一步分割. 因此针对 TransUNet 网络模型在分割咽后壁区域的上采样环节丢失部分语义信息的问题, 本文在解码器上采样部分前添加 FPN 结构, 每个分支都与编码器的不同层级特征相连, 由此更好地捕捉不同尺度的特征, 从而提高模型的准确性和鲁棒性.

2.2.4 改进后的网络结构

改进后的 TransUNet 如图 6 所示. 首先通过 CNN 提取模块获取特征图, 特征图经过深度可分离卷积模块局部特征提取和压缩, 获得低级别的图像特征, 有效地减少参数数量, 提升运算速度, 并且保持较好的特征表达能力. 接着经过每一次下采样后, 输入到 CBAM 模块进行注意力调整, 提高模型对重要特征的关注度, 对提取到的特征图进行序列化操作再传入 Transformer 层中, 将序列化后的特征图重新组合为完整的特征图, 并与解码部分进行跨层连接. 将 FPN 模块生成的多尺度特征金字塔作为输入, 每个分支都与编码器的不同层级特征相连, 对每个分支的特征图进行上采样, 以使其与其他分支的特征图具有相同的尺寸, 再利用横向连接合并所有分支的特征图, 并使用卷积层进行特征融合, 将融合后的特征图输入到解码器中进行进一步的上采样操作, 最后经过提高分辨率将特征图和逐步还原到原始图像尺寸进行分割.

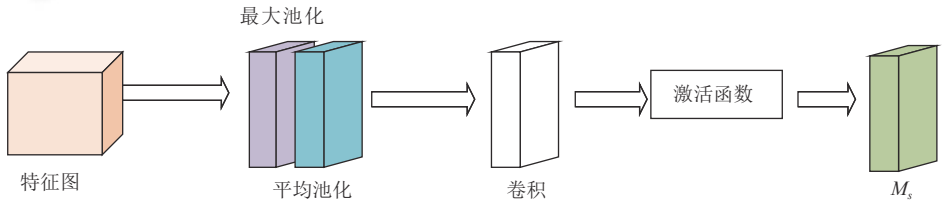


图 5 空间注意力模块

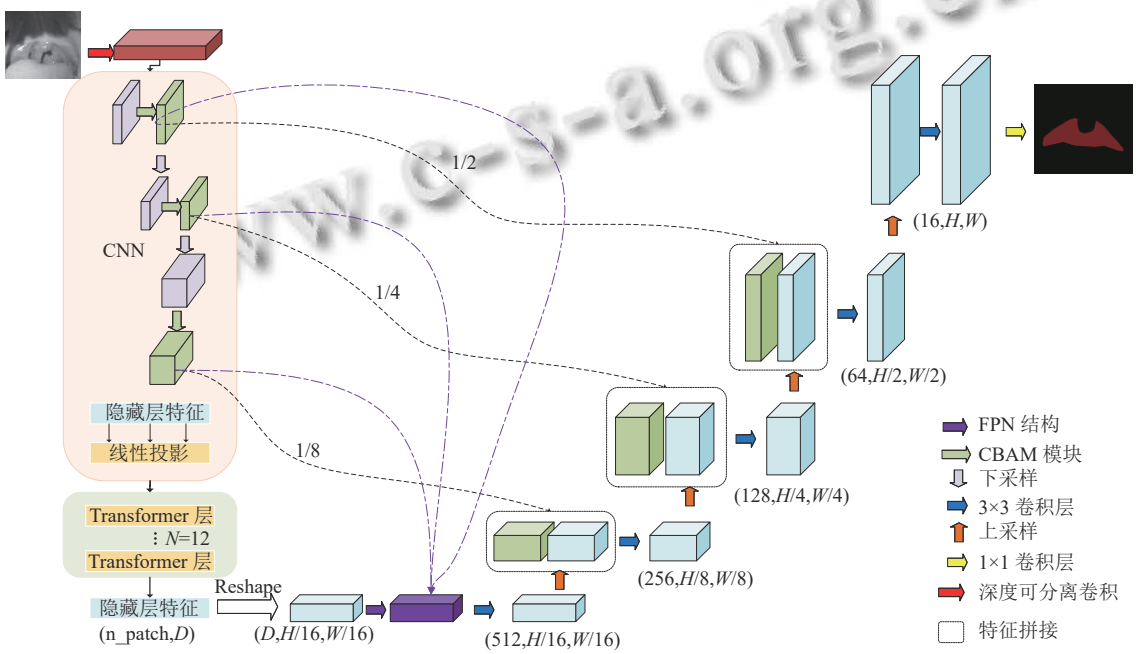


图 6 TransUNet 网络结构图

3 实验环境与数据集制作

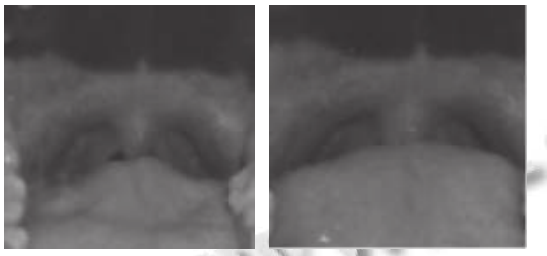
3.1 实验环境

本文所有训练测试都在如下环境进行: 实验平台是 PyCharm 2021, 编程语言是 Python 3.7, 操作系统是 64 位的 Windows 11, 其 CPU 是 i7-13700H, 显卡是 NVIDIA GeForce RTX 4060, 深度学习框架则是 PyTorch 1.4.

3.2 数据集获取与制作

3.2.1 数据集获取

训练目标检测模型以识别分割口腔内部咽后壁轮廓, 首要任务是解决数据集的问题. 目前相关咽部图像分割领域的研究相对较少, 公开的标准数据集几乎没有可用资源. 此外, 还需要考虑实际光线环境、噪声等因素对识别结果的影响由于悬雍垂、扁桃体等咽部组织都是软体, 很容易受气流和外部干扰等客观因素影响, 同一个人在同一姿势下的咽部 M 区域形状都可能会有所差异, 并且悬雍垂、舌头比较灵活, 不同的表情乃至呼吸都可能会改变悬雍垂和舌头的位置, 所以在制作数据集的时候需要考虑到采集不同姿势、不同表情下等状态下咽部的图像, 在拍摄过程中加入了拍摄角度、光线和咽喉变化等干扰因素以增加分割网络的抗干扰能力. 图 7 为同一名受试者在同一姿势下的咽喉变化对咽后壁采集图像的影响.



(a) 咽喉变化前 (b) 咽喉变化后

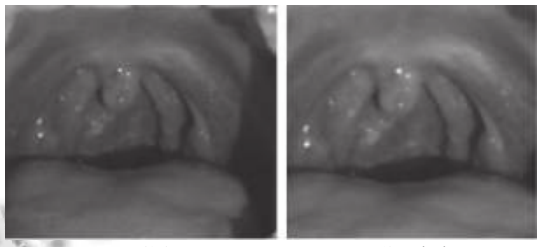
图 7 咽喉变化对咽后壁采集图像的影响

3.2.2 数据集制作

本实验利用 Intel RealSense D435i 深度相机共采集了 1 000 张面部口腔咽后壁数据图像, 用于构建训练深度学习目标检测与识别模型的训练集和测试集. 并采用亮度增强、镜像变换、对比度调整等方式扩充数据集, 最终获得 1 600 张咽后壁图像, 按照比例将数据集随机划分为训练集和测试集, 其中训练集有 1 300 张图像, 测试集有 300 张图像.

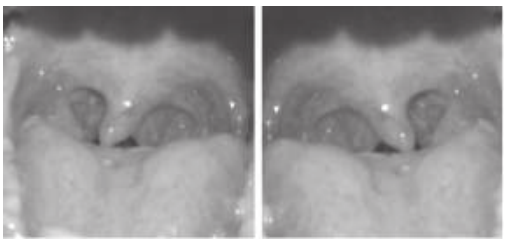
在拍摄图像过程中, 会出现光源被遮挡或者其他因素导致的咽后壁环境黑暗导致无法看清的情况, 在数据预处理过程中, 采用亮度增强来处理数据集. 图 8 是亮度增强了 20% 后的效果.

采用镜像变化处理扩增数据集. 在数据集里随机选取 200 张图片进行镜像变换. 图 9 是进行水平翻转之后的效果.



(a) 原图 (b) 增强亮度后

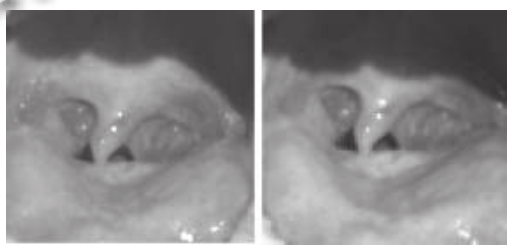
图 8 咽后壁图像亮度增强



(a) 原图 (b) 水平翻转后

图 9 咽后壁图像水平翻转

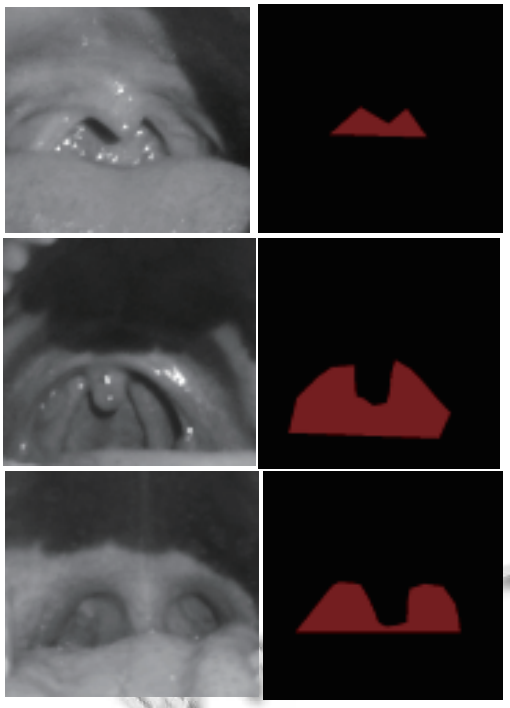
由于口腔内部存在粘液会导致镜面反射, 悬雍垂和口腔上颚等部位会在图片中显示过亮, 降低图像质量, 对图像分割效果产生影响. 这里采用对比度调整处理数据集, 图 10 是将对比度增强 30% 之后的效果.



(a) 原图 (b) 增强对比度后

图 10 咽后壁图像增强对比度

将处理完成的 1 600 张图片用图像标记工具 Labelme 对口腔咽后壁 M 区域进行人工标记, 将采样目标 M 区域从背景中划分出来, 以便于网络模型进行学习, 保存的数据为 png 格式, 且图像大小统一为大于 1270×720. 如图 11 所示, 左边为咽后壁原图像, 右边为标签图像.



(a) 原图 (b) 标签图像

图 11 制作标签图像

4 实验结果与分析

4.1 评价指标

为了评估 BM-TransUNet 在口腔咽后壁图像中分割 M 形状的可行性, 选取真正例、假正例、真负例、假负例 4 个指标对像素点进行统计. 真正例 (true positives, *TP*) 表示正确识别出的 M 区域的像素点数量; 假正例 (false positives, *FP*) 表示错误识别出的 M 区域的像素点数量; 真负例 (true negatives, *TN*) 表示被正确划分的口腔背景像素点的数量; 假负例 (false negatives, *FN*) 表示被错误划分的口腔背景像素点的数量. 并采用图像分割中通用的 4 种评价指标来进行效果分析: 精确率 (*Precision*) 、召回率 (*Recall*) 、*Dice* 系数和交并比 (intersection over union, *IoU*). 公式见式 (6)–式 (9):

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

$$Dice = \frac{2\times TP}{(TP+FN)+(TP+FP)} \tag{8}$$

$$IoU = \frac{TP}{TP+FP+FN} \tag{9}$$

其中, *Precision* 是指被正确预测为 M 区域的像素总数占被预测为 M 区域的像素总数的比例; *Recall* 是指模型正确识别出为正类的样本的数量占总的正类样本数量的比值; *Dice* 系数用来衡量预测结果与真实结果之间的相似程度; *IoU* 表示分割结果与真实标签间的重叠程度.

4.2 损失函数

损失函数采用交叉熵损失函数 (cross entropy loss, $Loss_{ce}$) 和 *Dice* 损失函数 ($Loss_{Dce}$) 相结合的方式, 来衡量模型预测值与真实值之间的差距以优化网络性能. 交叉熵损失函数可体现真实概率分布与预测概率分布之间的差异, 即所需部分和背景的不平衡问题;而 *Dice* 损失函数对模型的边缘预测进行更好的优化, 使预测到的咽后壁形状更符合预期效果. 两个损失函数的公式见式 (10) 和式 (11):

$$Loss_{ce} = -\sum_{i\in N}\sum_{l\in L} y_i \log\hat{y}_i \tag{10}$$

$$Loss_{Dce} = 1 - \frac{2\sum_{i=0}^{N} y_i\hat{y}_i}{\sum_{i=0}^{N}(y_i+\hat{y}_i)} \tag{11}$$

4.3 实验分析

为了验证 BM-TransUNet 网络模型的分割性能, 本文选取了 FCN、SegNet、DeepLabV2 、U-Net、Swin-UNet 等经典网络与原始的 TransUNet 网络模型进行对比, 并分析 BM-TransUNet 模型结构分割咽后壁的有效性. 将初始学习率 *Lr* 设置为 0.01, 并随着训练轮数的增加而自动调整逐渐降低, 如图 12 所示. 随着训练轮数的增多, 交叉熵损失函数 $Loss_{ce}$ 也逐渐降低趋于平稳, 如图 13 所示.

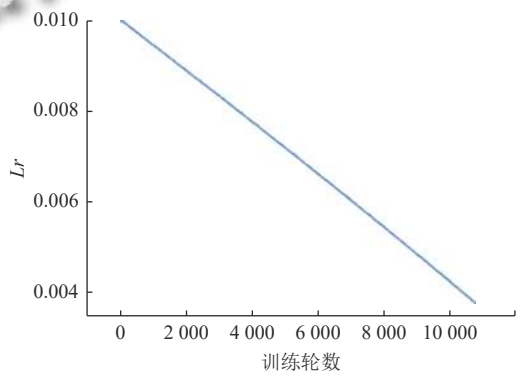


图 12 学习率变化曲线

表 1 是不同算法模型训练后的 *Loss* 值对比, 可以看出, 训练结束时 BM-TransUNet 的 *Loss* 值为 25.58%, 相较于其他模型来说损失值更少, 预测结果更接近真实值, 模型性能更稳定.

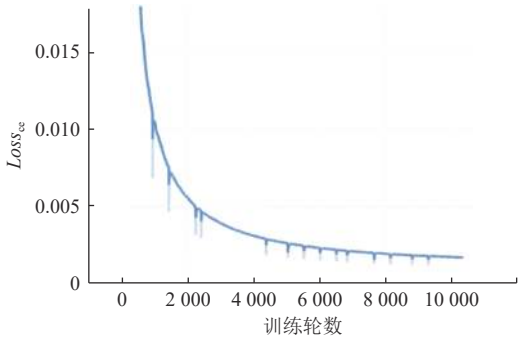


图 13 $Loss_{ce}$ 变化曲线

表 1 不同模型 *Loss* 值对比 (%)

算法模型	*Loss*
FCN	27.12
SegNet	26.95
DeepLabV2	26.09
U-Net	26.55
Swin-UNet	25.96
TransUNet	26.12
BM-TransUNet	25.58

4.3.1 算法模型对比实验

选取多种算法模型用于训练本文自制口腔咽后壁数据集, 训练后的 *Precision*、*Recall*、*Dice*、*IoU* 指标如表 2 所示.

从表 2 中可以看出, 与原始 TransUNet 网络相比, 优化后的 TransUNet 网络的检测精度更高, *Precision* 和 *Dice* 分别提升了 4.77% 和 2.01%, 并且在交并比 *IoU* 上与其他主流图像分割网络相比分别提升 9.72%、8.63%、2.58%、6.31%、1.06%、2.49%. 对比实验结果表明, 本文针对咽后壁 M 区域识别的模型优化是有效的, BM-TransUNet 网络的提取特征的能力更突出, 提取的 M 区域轮廓边界信息相比于其他边界信息更清晰连续, 总体精度较高.

表 2 不同模型指标结果对比 (%)

算法模型	*Precision*	*Recall*	*Dice*	*IoU*
FCN	76.56	77.69	79.62	80.13
SegNet	84.61	80.23	86.35	81.22
DeepLabV2	86.16	82.74	86.45	87.27
U-Net	83.21	85.39	84.43	83.54
Swin-UNet	87.39	86.71	90.60	88.79
TransUNet	88.84	92.63	88.75	87.36
BM-TransUNet	93.61	93.15	90.76	89.85

图 14 中: (a) 为原始图像, (b) 为标签图像, (c) 为 FCN 分割图像, (d) 为 SegNet 分割图像, (e) 为 DeepLabV2 分割图像, (f) 为 U-Net 分割图像, (g) 为 Swin-UNet 分割图像, (h) 为 TransUNet 分割图像, (i) 为 BM-TransUNet 分割图像. 从部分可视化分割效果对比图可得: 其他算法对于咽后壁 M 区域的分割情况出现了各种问题, 比如边缘锯齿状模糊、区域大小不同等问题. BM-TransUNet 不仅能够很好地改善上述问题, 而且能够有效提取到图像中的 M 区域的全局细节特征, 并保持信息的完整性, 从而实现了较优秀的分割效果.

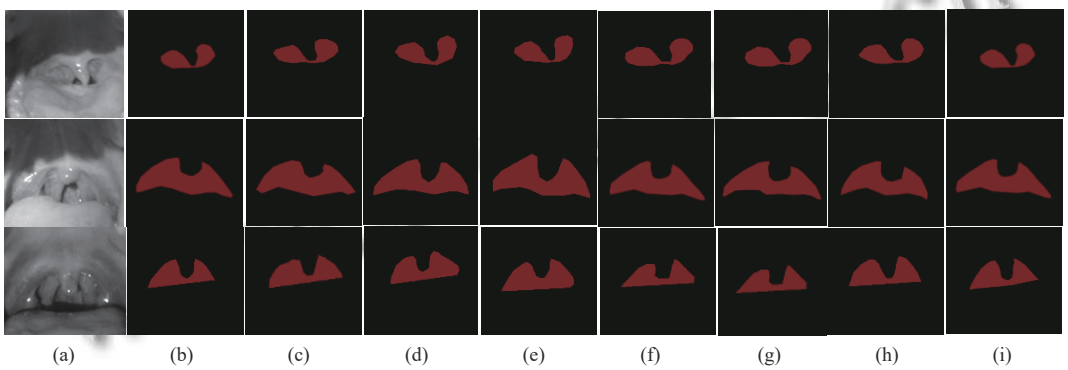


图 14 不同网络分割图像结果对比

4.3.2 CBAM 模块控制变量实验

基于 TransUNet 网络结构, 引入 CBAM 模块后进行模型训练, 并与原始网络结构进行控制变量对比, 训练后结果如表 3, 可以发现: *Precision* 指标提高了 3.81 个百分点, *IoU* 指标提升了 1.39 个百分点. 可以看出带有 CBAM 模块的模型在本文数据集上的准确率略高于不带 CBAM 模块的模型, 实验证明 CBAM 模块可以提升模型的泛化能力, 减少过拟合的风险. 此外, 带 CBAM 模块的模型 *IoU* 指标的提升表示预测结果和真实分割匹配, 重叠度高, 说明模型对目标的分割效果变得更好.

4.3.3 消融实验

为了验证以 TransUNet 为基础所作网络改进的有效性, 通过分别添加深度可分离卷积模块 DSC 和多尺度特征融合模块 FPN, 并以相同数据集来进行消融实

验, 同时用 *Precision*、*Dice*、*IoU* 来进行效果评价, 结果如表 4 所示. 实验表明, 当算法中只添加 DSC 模块时, *Precision* 相比于原始网络提高了 0.94%, 此外, *IoU* 也上升了 0.53%, 有一定的分割效果; 同时, 在仅有多尺度特征融合模块时, 3 个指标相比于 TransUNet 网络分别增长了 3.4%、0.67%、1.25%, 均有显著提升. 由此可见本文所作改进是有效的, 所添加模块对算法优化改进均有正向效果.

表 3 CBAM 模块控制变量实验结果 (%)

网络结构	*Precision*	*IoU*
TransUNet	88.84	87.36
TransUNet+CBAM	92.65	88.75

表 4 添加不同模块对分割效果的影响 (%)

方法	*Precision*	*Dice*	*IoU*
TransUNet	88.84	88.75	87.36
TransUNet+DSC	89.78	88.96	87.89
TransUNet+FPN	92.24	89.42	88.61
BM-TransUNet	93.61	90.76	89.85

5 结束语

本文提出了一种基于 TransUNet 的改进算法 BM-TransUNet, 在算法编码器部分加入了深度可分离卷积模块和注意力机制模块 CBAM, 可以提升口腔咽后壁 M 区域的分割精度, 更完整地提取到了 M 区域的全局信息特征, 更好学习图像精细特征, 弥补细节损失, 同时在编码器与解码器间添加了多尺度特征融合模块 FPN, 对来自编码器的特征进行处理并合并所有分支的特征图, 经卷积操作后输入到解码器中进行进一步上采样操作, 以恢复目标分割结果. 利用自制的咽后壁数据集, 将改进后的网络与 FCN、SegNet 、UNet、TransUNet 等几种网络结构对数据集进行训练并对比, 实验结果表明: BM-TransUNet 算法 *Precision*、*Recall*、*Dice* 系数和 *IoU* 指标更优于其他算法, 并通过控制变量实验和消融实验证明了 BM-TransUNet 的有效性.

参考文献

1 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436–444. [doi: 10.1038/nature14539]

2 Gu ZW, Cheng J, Fu HZ, *et al*. CE-Net: Context encoder network for 2D medical image segmentation. IEEE Transactions on Medical Imaging, 2019, 38(10): 2281–2292. [doi: 10.1109/TMI.2019.2903562]

3 Li PX, Zhao HC. Monocular 3D detection with geometric constraint embedding and semi-supervised training. IEEE Robotics and Automation Letters, 2021, 6(3): 5565–5572. [doi: 10.1109/LRA.2021.3061343]

4 Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640–651. [doi: 10.1109/TPAMI.2016.2572683]

5 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Proceedings of the 18th International Conference on Medical Image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234–241.

6 Park B, Park H, Lee SM, *et al*. Lung segmentation on HRCT and volumetric CT for diffuse interstitial lung disease using deep convolutional neural networks. Journal of Digital Imaging, 2019, 32(6): 1019–1026. [doi: 10.1007/s10278-019-00254-8]

7 金鹭, 张寿明. 基于 U-Net 网络改进算法的视网膜血管分割研究. 光电子·激光, 2022, 33(8): 887–896.

8 马豪, 刘彦, 张俊然. 基于模型压缩与重构 U-net 的胰腺分割. 计算机工程与设计, 2022, 43(7): 1998–2006.

9 VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017. 6000–6010.

10 Chen JN, Lu YY, Yu QH, *et al*. TransUNet: Transformers make strong encoders for medical image segmentation. arXiv:2102.04306, 2021.

11 韩文龙. 基于 TransUNet 的微小水体遥感信息提取算法研究 [硕士学位论文]. 廊坊: 北华航天工业学院, 2022.

12 Chen JY, Chen JN, Zhou ZW, *et al*. MT-TransUNet: Mediating multi-task tokens in Transformers for skin lesion segmentation and classification. arXiv:2112.01767, 2021.

13 Wang N, Lin SH, Li XX, *et al*. MISSU: 3D medical image segmentation via self-distilling TransUNet. IEEE Transactions on Medical Imaging, 2023, 42(9): 2740–2750. [doi: 10.1109/TMI.2023.3264433]

14 Chang CY, Lin TK, Lin CW, *et al*. Application of TransUNet for segmenting lung mass from chest X-ray image. Proceedings of the 2022 IEEE International Conference on Consumer Electronics. Taipei: IEEE, 2022. 175–176.

15 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. Proceedings of the 9th International Conference on Learning Representations. ICLR, 2021.

16 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.

17 Carion N, Massa F, Synnaeve G, *et al*. End-to-end object detection with transformers. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 213–229.

(校对责编: 孙君艳)