DOI:10.11721/cqnuj20220615

# 一种 CNN-Transformer 网络在皮肤镜图像分割上的应用*

董玉民，卫力行
(重庆师范大学 计算机与信息科学学院，重庆 401331)

**摘要**:【目的】针对皮肤病变图像存在皮损形状不规则、边界模糊以及毛发伪影等问题，提出了一种将 CNN 和 Transformer 相结合的图像分割算法。【方法】首先对皮肤病变图像进行去毛发预处理，减少毛发噪声对结果的影响，然后构建 CNN 和 Transformer 结合的分割模型，采用 Resnet 作为特征提取主干网络，将提取到的特征图序列作为 Transformer 的输入，在 Transformer 中加入了新的结构边界注意门以提取足够的局部细节来处理模糊边界，最后采用 DenseASPP 模块增强特征表示和处理多尺度信息，并且提出一种改进了的损失函数，以便在计算损失函数的同时使得模型能关注边界区域部分。【结果】提出的算法在 ISIC2017 数据集上的 Dice 指数值以及 Jaccard 指数值分别为 0.854 534 和 0.767 901，在 ISIC2018 数据集上的 Dice 指数值以及 Jaccard 指数值分别为 0.908 548 和 0.843 689，与其他算法相比提出的算法对图像的分割效果相对较好。【结论】实验结果证明了所提算法在皮肤病变图像上进行的图像分割是有效的。

**关键词**:CNN;Transformer;DenseASPP;皮肤病变分割
**中图分类号**:TP301.6　　**文献标志码**:A　　**文章编号**:1672-6693(2023)02-0126-09

近些年来，基于数据驱动的深度学习方法被广泛应用到医学图像分割领域，该方法的优势在于依靠数据驱动自动学习图像特征从而在图像识别等方面表现出良好的性能。目前，为了实现精确分割，大多数研究人员利用 CNN 来设计基于深度学习的方法，其中最典型是 Long 等人[1]提出的 FCN——该算法采用端到端、像素到像素的方式超过了当时的最先进水平。尽管 FCN 对医学图像处理工作提供了很大的帮助，但是仍存在 FCN 训练比较麻烦，且得到的结果不够精细等问题。为了解决 FCN 的不足，作为生物医学图像分割的经典框架，U-Net[2]综合上述方法通过在编码器和解码器层之间添加快捷方式实现了高性能分割。由于 U-Net 在医学图像分割领域中的出色表现，因此现阶段医学图像分割算法多由 U-Net 演化而来。在皮肤镜图像分割中，Wen[3]采用了具有 4 个卷积分支的 inception 模块[4]，这种模块具备非线性特征学习能力，可以应用于层数较少的 FCN 模型，此外还利用了条件随机场[5]来细化处理分割的结果。Thao 等人[6]通过全卷积-反卷积网络将皮肤肿瘤与周围皮肤从图像中分割开来。Yu 等人[7]引入深度残差网络[8]用于皮肤病变的自动分割，他们通过将多个残差块堆叠起来，提高了模型的表达能力。上面提到的方法在关于皮肤的图像分割方面也许可以取得比较好的结果，但是这些方法均存在只能产生有限感受野的问题，即无法捕获全局依赖性。最近，有学者提出通过自我注意机制将图像划分成一系列小块从而获取全局视野，最典型的 TransUnet 采用 CNN 和 Transformer 的混合架构在多器官的图像分割方面表现良好[9]。Transformer 被认为是一种很有前途的全局上下文建模工具，它采用了一种强大的全局注意机制，但它在分割任务中的主要缺点之一是无法有效地提取足够的局部细节来处理模糊边界[10]。

本文提出了一种新的算法：首先使用预处理步骤，先去除一些皮肤镜图像本身噪声影响，比如毛发噪声；其次采用一种新的 Transformer 和 CNN 的混合架构，在获取足够的全局视野的同时也能够提取足够的局部细节来处理模糊边界；然后采用 DenseASPP[11]来增强特征表示和处理多尺度信息；最后采用一种新的损失函数以提高对边界的关注度。本文通过与不同的模型之间对图像分割的效果对比证明了所提算法的有效性。

## 1 相关问题描述

皮肤癌是所有癌症中比较常见一种，导致这种病症发生的最根本原因是人体长期暴露在紫外线下产生了一

* **收稿日期**:2022-04-11　**修回日期**:2022-06-01　**网络出版时间**:2022-12-09T09:52
**资助项目**:国家自然科学基金面上项目(No. 61772295; No. 61572270; No. 61173056);重庆市科学技术局项目(No. cstc2021jsyj-yzysbA0042);重庆市教育委员会科技攻关计划项目(No. KJZD-M202000501);重庆市技术创新与应用开发专项普通项目(No. cstc2020jscx-lyjsA0063)
**第一作者简介**:董玉民，男，教授，博士，研究方向为医学图像处理、量子信息、人工智能等，E-mail: dym@cqnu.edu.cn
**网络出版地址**:https://kns.cnki.net/kcms/detail//50.1165.n.20221207.1824.017.html

种被称为黑色素瘤的肿瘤[12]。黑色素瘤是一种发展迅速、死亡率较高的恶性肿瘤。美国皮肤癌协会的统计数据报告显示，黑色素瘤在全球最常见的癌症问题中排名第 19 位[13]。该统计报告还显示：2021 年美国男性中大约出现 62 260 例黑色素瘤病例，占总体恶性肿瘤病例的 6%；女性中大约出现 43 850 例皮肤恶性肿瘤病例，占总体恶性肿瘤数的 5%。当前人类社会中黑色素瘤的发病率逐年增加，且在肿瘤发生转移后，2 年生存率不足 15%，5 年生存率不足 5%，因而对于肿瘤的早期发现与治疗对阻止黑色素瘤的扩散有着极为重要的意义[14]。

目前对该疾病的诊断在临床上主要采用采样皮损皮肤镜图像的方式。这是一种无创的成像技术，然而皮肤科医师通过肉眼观察皮肤镜图像并做出诊断往往费时费力，因此计算机辅助系统被广泛应用于检测黑色素瘤，在用计算机辅助系统进行检测的过程中，病灶的区域分割是详细分析病变结构的最重要过程。但是黑色素瘤的病变区域存在 3 个问题：1) 周围皮肤与病变区域对比度较低(封二彩图 1a)；2) 皮肤镜图像中存在纹理和毛发等噪声(封二彩图 1b)；3) 病变区域形状不规则，且边界模糊(封二彩图 1c)。这 3 个问题使得分割任务非常具有挑战性，临床上多采用耗时耗力的手工方式来分割图像，因此急需引入客观可靠的自动分割技术来代替手工方式，对皮肤的病灶区域进行准确的分割，从而提高效率并避免主观因素的影响，这对黑色素瘤的早期诊断和准确切除有着重要意义[15]。

## 2　CNN-Transformer 结构解决方法

### 2.1　模型

本文模型融合了典型的 CNN 和 Transformer 架构，主要包含以下几个部分：CNN 架构主干特征提取网络用于提取特征；Transformer 中的 self-attention 考虑全局来粗略定位病变区域，边界注意门(Boundary gate，BG)来捕获更多的局部细节；采用 DenseASPP 模块来获取更多的多尺度上下文信息。整个模型总体架构如封三彩图 2。

#### 2.1.1　Resnet 主干特征提取网络(CNN 架构)

通过 VGG[16]以及 GoogleNet[17]的实验证明层次更深的网络具有更强的非线性表达能力，但是随着网络层次的加深往往会出现梯度消失等问题。Resnet 网络[18]的提出有效地克服了加深网络结构加深时梯度消失的问题，可以得到更好的预测效果。Resnet 提出了一种残差模块，通过堆叠残差模块可以构建任意深度的神经网络而不会出现退化的现象，同时采用批归一化方法来对抗梯度消失的问题，该批归一化方法降低了网络训练过程对于权重初始化的依赖，并提出了一种针对 ReLu 激活函数的初始化方法。残差模块的计算公式为：$H(x)=F(x)+x$。式中：$F(x)$指的是卷积层学习的变换，$H(x)$表示残差结构的输出，残差模块相当于一个恒等映射，加上 $x$ 保证学习到的内容至少是输入层的内容，并且回传的梯度不丢失，$F(x)$学习到的是感兴趣的特征，$F(x)+x$ 得到的是对输入层 $x$ 内容的增强(输入层内容中包含感兴趣的特征)，以保证原来的信息不丢失。

#### 2.1.2　Transformer 架构

目前 Transformer 在自然语言处理领域中应用广泛，最近也逐渐运用于图像处理领域。Transformer 相对于 RNN、CNN 而言具有很多优势，其中最重要的优势就在于并行计算以及全局视野。相较于 RNN 的序列模型而言，采用并行计算可以大大提高效率；而相较于 CNN 来说，可以不需要通过堆叠卷积核的方式就获得更大的感受野。本文在传统的 Transformer 结构的基础之上增加了一个 BG 边界门模块，具体结构如封三彩图 3 所示。

编码器的主要构成部分如下：

1) 特征图序列与位置编码。在进行卷积神经网络提取特征之后，会产生很多的高级语义特征，它们由特征响应图反应出来，现在要将特征图序列进行位置编码作为 Transformer 的输入，循环神经网络采用顺序迭代的方式处理数据，Transformer 模型在循环神经网络的基础上做出了改进，采用并行计算的方式大大提高了效率，但是在这个过程中首先必须进行位置编码以提供每个特征块的位置信息，才能识别特征块之间的顺序关系。位置编码后的维度为[max_sequence_length，batchsize，embedding_dimension]，其中 max_sequence_length 为特征图被分割成的 patch 的数目，embedding_dimension 表示为了将 patch 变平并通过可训练的线性投影后的嵌入维度，batchsize 为批处理个数。在这里 patch 的大小是 32×32(可参考封三彩图 2)，将每个 patch 展平成向量的形式，向量维度也就等于特征图的大小 1 024(即 32×32)。

2) 多头注意力机制。编码器由 $n$ 个堆叠的编码器层组成，用于在整个皮肤镜图像中捕获远程上下文。编码器遵循典型 Transformer 的设计，它里面的每一层大都由一个多头注意力模块(multi-head attention，MSA)和一

个前馈网络(feed-forward network,FFN)组成。其中多头注意力模块用于过滤非语义信息,进一步精细空间的恢复,让网络粗略定位到关键信息部分。多头注意力模块的主要目标是将深层次的特征图中的每个元素相互连接起来,从而获得全局视野。其中 self-attention 模块是多头注意力模块中的重要组成部分。一个 self-attention 模块接受 3 个输入,分别是查询矩阵 $\boldsymbol{Q}$、键矩阵 $\boldsymbol{K}$ 和值矩阵 $\boldsymbol{V}$。self-attention 模块的计算公式为:

$$\begin{cases}\boldsymbol{Q} = \text{Linear bmod}_{\boldsymbol{Q}}(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{W}_{\boldsymbol{Q}},\\ \boldsymbol{K} = \text{Linear bmod}_{\boldsymbol{K}}(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{W}_{\boldsymbol{K}},\\ \boldsymbol{V} = \text{Linear bmod}_{\boldsymbol{V}}(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{W}_{\boldsymbol{V}}。\end{cases}$$

式中:$\boldsymbol{X}$ 表示进行位置编码后的输入,将 $\boldsymbol{X}$ 与 $\boldsymbol{W}_{\boldsymbol{Q}}$,$\boldsymbol{W}_{\boldsymbol{K}}$,$\boldsymbol{W}_{\boldsymbol{V}}$ 相乘便得到 $\boldsymbol{Q}$,$\boldsymbol{K}$,$\boldsymbol{V}$。这里的 $\boldsymbol{W}_{\boldsymbol{Q}}$,$\boldsymbol{W}_{\boldsymbol{K}}$,$\boldsymbol{W}_{\boldsymbol{V}}$ 均表示权重参数矩阵,它们都是随机初始化的,通过模型学习而不断优化。

接下来将 $\boldsymbol{Q}$ 和 $\boldsymbol{K}^{\mathrm{T}}$ 相乘作为相似度系数,然后除以 $\sqrt{d_k}$,经过 softmax 以后再乘以 $\boldsymbol{V}$ 得到输出,计算公式为:$\boldsymbol{X}_{\text{attention}} = f_{\text{softmax}}[\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}})/\sqrt{d_k}]\boldsymbol{V}$。

3) 残差连接和分层归一化(layer normalization)。接下来开始进行残差连接以及分层归一化部分,在这里首先得到了上一步经过 self-attention 加权和的结果,这里记作 $\boldsymbol{X}_{\text{attention}}$,然后将它加起来做残差连接,得到 $\boldsymbol{X}_{\text{output}} = \boldsymbol{X}_{\text{embedding}} + \boldsymbol{X}_{\text{attention}}$。式中的 $\boldsymbol{X}_{\text{embedding}}$ 是特征图位置编码后得到的编码矩阵,以提供每个特征图的位置信息,以此来识别特征图之间的顺序关系。

在残差连接后,通过分层归一化将神经网络的隐藏层归一化为标准正态分布,以加快训练和收敛速度。根据下列公式来求 $\boldsymbol{X}_{\text{output}}$ 归一化后的值:

$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_{ij},\sigma_j^2 = \frac{1}{m}\sum_{i=1}^{m}(x_{ij}-\mu_j)^2,\ \boldsymbol{X}_{\text{output}} = f_{\text{LayerNorm}}(\boldsymbol{X}_{\text{output}}) = \left[(x_{ij}-\mu_j)/\sqrt{\sigma_j^2+\varepsilon}\right]_{ij}。$$

这里为了防止分母为 0,在归一化的计算中也即 $\boldsymbol{X}_{\text{output}}$ 计算式的分母里添加了一个正数 $\varepsilon$。

4) 前向传播(feed forward)。前向传播其实就是两层线性映射并用激活函数来激活,如激活函数 ReLu。前向传播的计算公式为:$\boldsymbol{X}_{\text{hidden}} = \text{Linear}(\text{ReLU}(\text{Linear}(\boldsymbol{X}_{\text{layerNorm}})))$。

5) 将前向传播后的结果再次残差连接与分层归一化,具体计算公式为:$\boldsymbol{X}_{\text{hidden}} = \boldsymbol{X}_{\text{attention}} + \boldsymbol{X}_{\text{hidden}}$,$\boldsymbol{X}_{\text{hidden}} = f_{\text{layerNorm}}(\boldsymbol{X}_{\text{hidden}})$。

6) BG。处理边界信息模块可以在处理边界模糊的病变时获得更大的能力,为此设计了 BG。在每个 Transformer 编码层的末端添加一个边界注意力,以细化处理后的特征。BG 的体系结构类似于传统的空间注意力机制[19](图 4)。

首先特征映射 $\boldsymbol{F}$ 分别生成两个新的特征映射 $\boldsymbol{F}_1$ 和 $\boldsymbol{F}_2$,这里的 $\boldsymbol{F}$,$\boldsymbol{F}_1$,$\boldsymbol{F}_2$ 的维度假设都用 $C\times H\times W$ 来表示。改变 $\boldsymbol{F}_1$ 及 $\boldsymbol{F}_2$ 的维度,由 $C\times H\times W$ 转变成 $C\times N$,这里 $N=H\times W$。然后将 $\boldsymbol{F}_1$ 与 $\boldsymbol{F}_2^{\mathrm{T}}$ 两个矩阵相乘,将运算的结果通过 softmax 便得到空间注意力的权重:

$$W^{i,j} = \left[\exp(\boldsymbol{F}_1^i \cdot \boldsymbol{F}_2^j)\right] \Big/ \left[\sum_{i=1}^{N}\sum_{j=1}^{N}\exp(\boldsymbol{F}_1^i \cdot \boldsymbol{F}_2^j)\right],$$

式中:$W^{i,j}$ 记录的是不同位置的权重,$i$,$j$ 分别表示行号和列号,$\boldsymbol{F}_1^i$ 和 $\boldsymbol{F}_2^j$ 分别表示 $\boldsymbol{F}_1$ 的 $i$ 行向量与 $\boldsymbol{F}_2$ 的 $j$ 列向量。

同时,将特征 $\boldsymbol{F}$ 输入到另一个卷积层,生成一个新的特征映射 $\boldsymbol{F}_3$,改变 $\boldsymbol{F}_3$ 的维度由 $C\times H\times W$ 转变成 $C\times N$,同样的 $N=H\times$

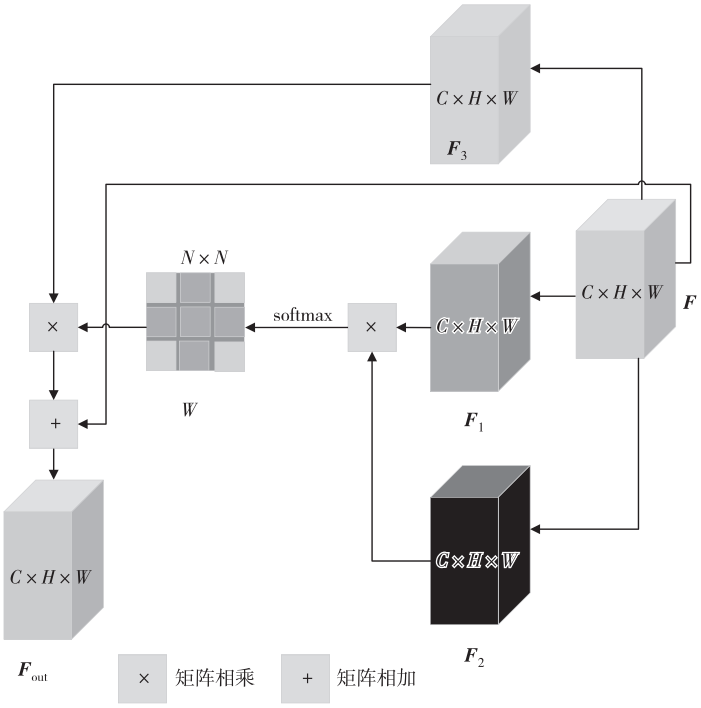



**图 4 边界门网络结构**

**Fig. 4 Boundary gate network structure**

$W$。然后执行 $\boldsymbol{F}_3$ 和 $\boldsymbol{W}^{\mathrm{T}}$ 之间的矩阵乘法运算，并将结果维度由 $C\times N$ 变换为 $C\times H\times W$。最后，用尺度参数 $\alpha$ 乘以结果并与 $\boldsymbol{F}$ 求和运算，得到最终的输出特征 $\boldsymbol{F}_{\mathrm{out}}$。这里的 $\alpha$ 初始化为 0，并通过学习逐渐变化。

和编码器一样，解码器的 3 个部分中的每个部分都有一个残差连接，后面接一个分层归一化操作。解码器的中间部件并不复杂，大部分在前文编码器的部分里已经介绍过了这里不再赘述。

### 2.1.3　DenseASPP

为了在多个尺度上分割病变，该模块将自我注意机制后的变换特征作为输入，旨在产生密集的预测以增强局部特征表示和处理多尺度病变信息。

研究表明[20]，空间金字塔池化(spatial pyramid pooling，SPP)结构对不同尺度的特征进行重采样加特征映射的感受野可以有效准确地分类任意大小的区域。为了增加感受野，同时不改变它的分辨率，Chen 等人[21]引入了空洞卷积，并且提出了空洞空间金字塔池化(atrous spatial pyramid pooling，ASPP)模块。该模块使用不同采样率的并行空洞卷积对给定的特征层重新采样，从而能够在多个尺度上捕捉目标。本文采用了在此基础上提出的 DenseASPP[11]模块，该模块通过一系列的空洞卷积，使得层数越深的神经元获得的感受野越大，而不受 ASPP 内核退化问题的困扰。同时，通过一系列的特征拼接，DenseASPP 最终生成的特征映射所覆盖的语义信息不仅尺度范围大，而且非常密集。DenseASPP 结构如图 5 所示。

图 5 中的输入为 $C\times H\times W$ 的特征图。通过不同的空洞率进行卷积可以获得不同感受野大小的结果图，将不同大小感受野的特征图进行拼接可以获得多尺度的上下文信息；同时，采用密集连接可以获得更多的信息，并且相当于多个子网络的集成，从而获得更好的性能。

### 2.1.4　辅助模块

在这里加入一种辅助模块以完成训练，通过辅助模块可以增强边界特征可以更好的处理模糊的病变边界，同时该设计还可以帮助加速训练网络模型，在相对较小的数据集上也能有效的学习到位置嵌入。

首先在原始标签图片上利用边缘检测算法生成边界点集 $k$，以边界点集 $k$ 中的每一个点为中心生成一个半径为 $r$ 的圆面区域(这里 $r$ 的大小设定为 15)，该圆区域的面积为 $S_1$，该圆区域与病变区域的交面积为 $S_2$，用 $p$ 表示圆区域中病变区域的比例($p=S_1/S_2$)，$p$ 的值越大(越小)则边界越不光滑(光滑)。将每个边界点都记下 $p$ 值，然后利用非最大值抑制对比每个点与相邻点的 $p$ 值以筛选出这些最具代表性的特征点。最后这些特征点的 2 维位置坐标($x$，$y$)映射到 1 维的位置，记为$\lfloor x/16\rfloor\times 16+\lfloor y/16\rfloor$，将特征点的标签设定为 1，其他位置设定为 0，也就生成了最终的辅助标签(在这里辅助标签以 npy 的文件格式存储)。

最后利用生成的辅助标签与 Transformer 中的编码器的输出进行对比，形成新的边界损失，使得模型更加关注于关键的边界特征点以便模型能够更好地处理模糊边界，同时也可以加快 Transformer 的训练。该辅助模块如图 6、图 7 所示。

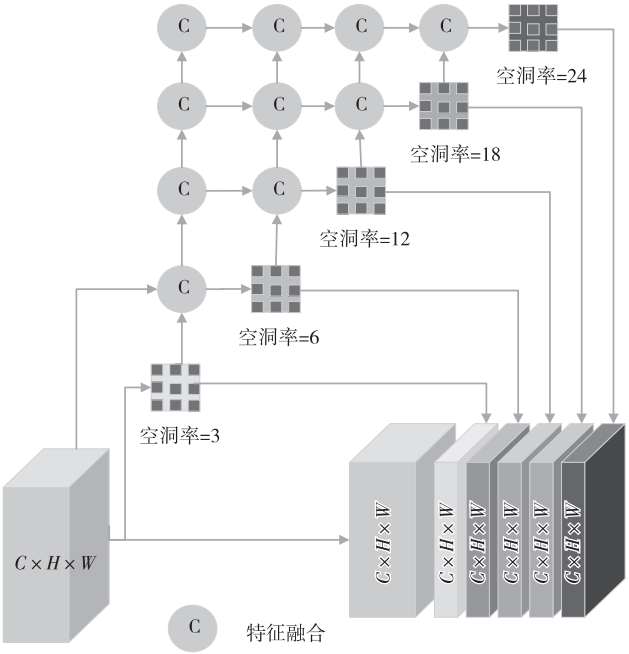



**图 5　DenseASPP 总体结构**

**Fig. 5　DenseASPP network structure**

## 2.2　损失函数

普通的损失函数例如交叉熵损失函数或者 BCE 损失函数也能取得较好的结果，但是在此次任务中不仅要定位到病变区域，还要求能精准地预测具体的轮廓从而不会漏掉部分细节。针对这一要求，本文改进了当前损失函数，目的就是在计算损失函数的时，使得模型能关注边界区域部分。

为了构造边界损失，这里用到了上文所涉及的辅助模块，利用生成的辅助标签与 Transformer 中的编码器的输出进行对比，形成新的边界损失，使得模型更加关注于关键的边界特征点以便

模型能够更好地处理模糊边界,同时也可以加快 Transformer 的训练。该边界损失或者辅助监督的计算过程如图 8 所示。

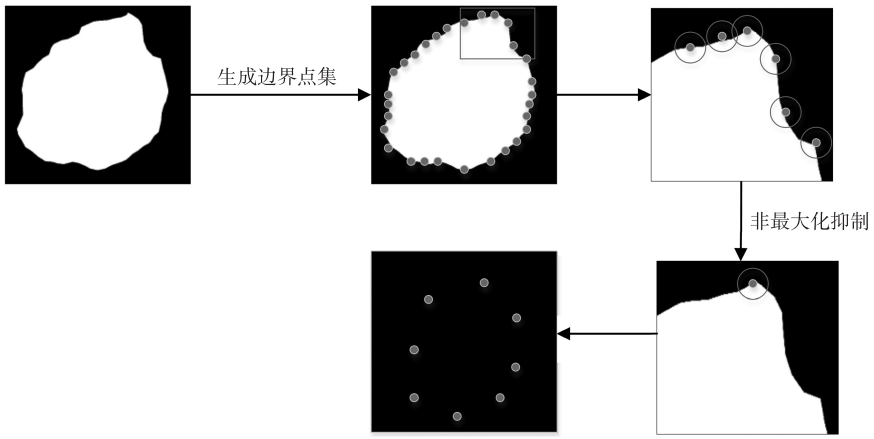



**图 6 生成辅助标签过程**

**Fig. 6 Generate a diagram of the auxiliary labeling process**

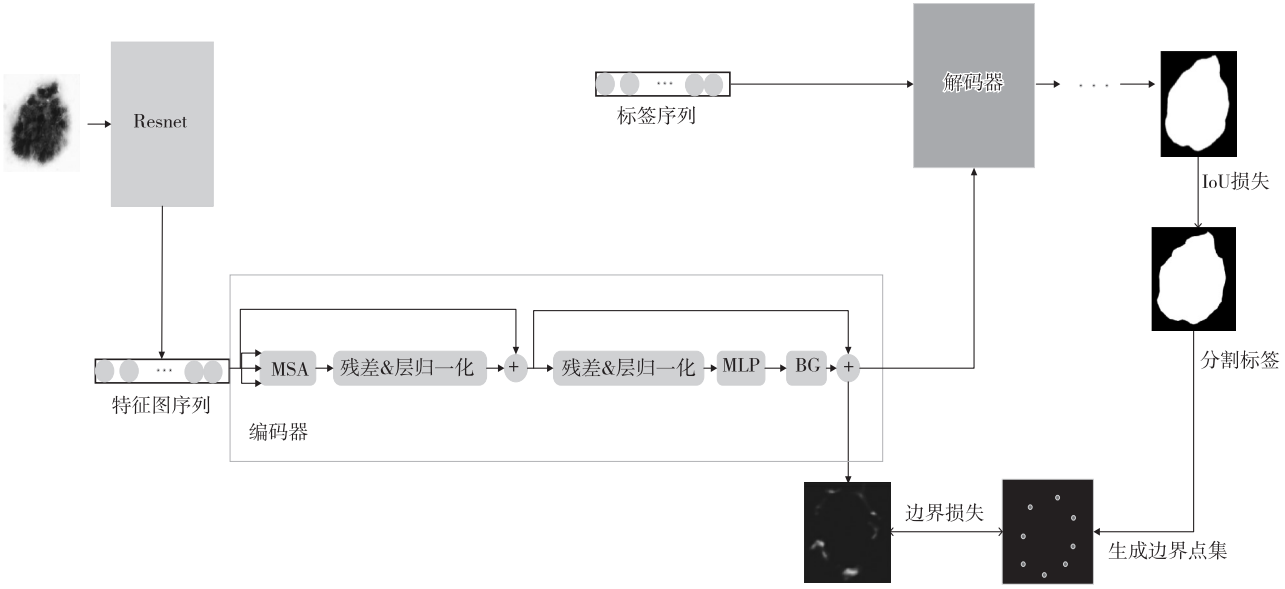



**图 7 辅助模块过程**

**Fig. 7 Auxiliary module process**

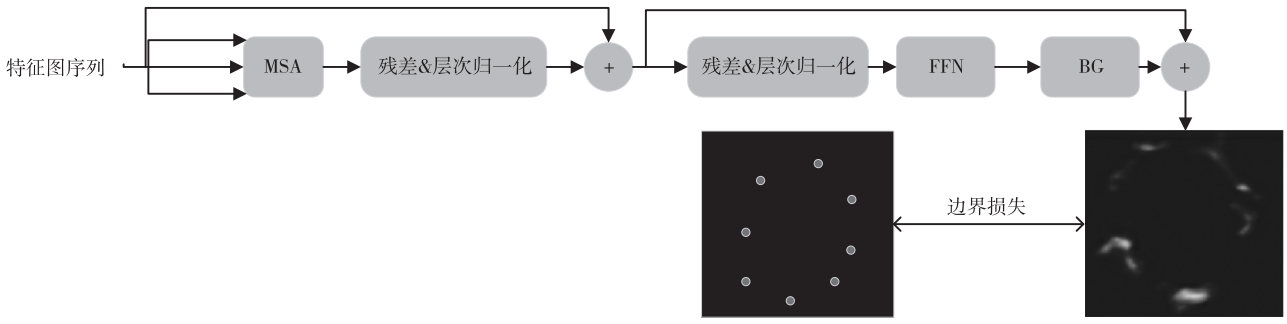



**图 8 边界损失或辅助监督计算过程**

**Fig. 8 Boundary loss or auxiliary supervision calculation process**

边界损失计算公式为:$f_{\mathrm{LoE}}=\sum_{n=1}^{k} f_{\mathrm{CeLoss}}^{n}(M_{\mathrm{pred}},M_{\mathrm{GT}})$。式中:$k$ 表示编码器的个数,$M_{\mathrm{pred}}$ 表示经过编码器的预测结果,$M_{\mathrm{GT}}$ 表示辅助模块生成的辅助标签,$f_{\mathrm{CeLoss}}$ 为多分类交叉熵损失函数,公式为:$f_{\mathrm{CeLoss}}=\frac{1}{N}\sum_{i} L_i=$

$-\frac{1}{N}\sum_{i}\sum_{c=1}^{M} y_{ic}\log(p_{ic})$。式中：$M$ 为类别数量，$y_{ic}$ 为符号函数(值为 0 或 1)，如果样本的真值等于 $c$ 取 1，否则取 0。$p_{ic}$ 表示观测样本 $i$ 属于类别 $c$ 的预测概率。

结合边界损失函数与 IoU 损失函数得到最终的复合损失函数为：

$$f_{\mathrm{LoA}}=f_{\mathrm{LIoU}}(T_{\mathrm{pred}},T_{\mathrm{GT}})+f_{\mathrm{LoE}}(M_{\mathrm{pred}},M_{\mathrm{GT}})。$$

这里的 IoU 损失函数 $f_{\mathrm{LIoU}}$ 为交并比损失函数，表示的就是真实病变区域和预测病变区域的交并比，即：$f_{\mathrm{LIoU}}=1-|T_{\mathrm{pred}}\cap T_{\mathrm{GT}}|/|T_{\mathrm{pred}}\cup T_{\mathrm{GT}}|$，式中：$T_{\mathrm{pred}}$ 表示最终网络预测的概率分割图，$T_{\mathrm{GT}}$ 表示标签图。

### 2.3　评价指标

在图像分割评价方法中，有监督评价方法是指比较算法预测结果以及真实结果来进行评价。其中真实结果是指真值图像(ground truth)或金标准(golden standard)。通过参考真值图像，可以让最终的评价更加的准确，这也是目前图像分割领域使用最多的一种评价方法。图像分割的有监督评价方法是基于分割图像与真值图像之间的相似度或差异度的，如果这两种图像之间的相似度越大或者差异度越小则代表着该分割算法越好。

对于真值图像和分割图像，包含真阳性、假阴性、假阳性、真阴性这几个概念。其中真阳性(true positive，TP)是指分割算法将目标像素正确地分类为目标；假阴性(false negative，FN)是分割算法将目标像素错误地分类为背景；假阳性(false positive，FP)是分割算法将背景像素错误地分类成目标；真阴性(true negative，TN)是分割算法将背景像素正确地分类为背景，具体如图 9 所示。

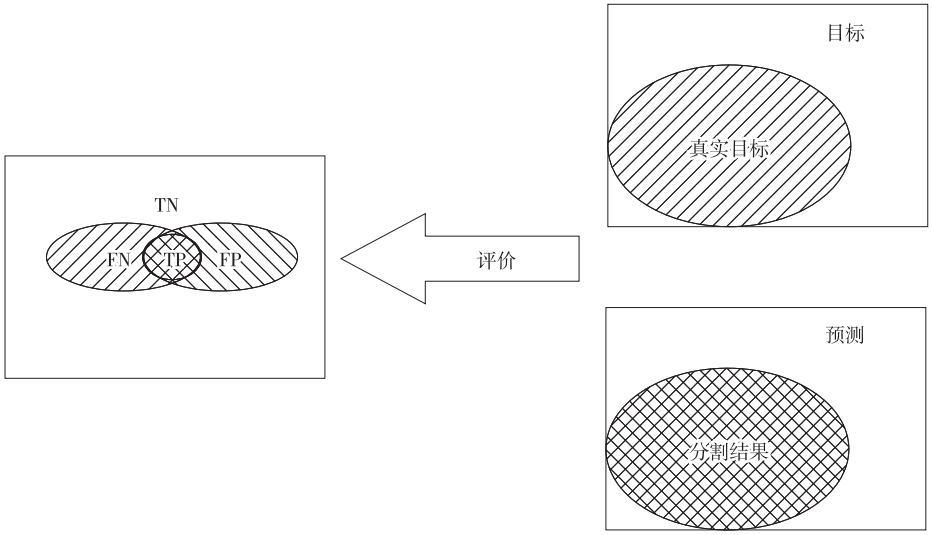



**图 9　真阳性、假阴性、假阳性和真阴性的定义**

**Fig. 9　Definition of true positive, false negative, false positive, and true negative**

几个常用的有监督评价指标如下：

1) 灵敏度(sensitivity)又称为查全率(recall)或真阳性率，定义为：$I_{\mathrm{Sensitive}}=V_{\mathrm{TP}}/(V_{\mathrm{TP}}+V_{\mathrm{FN}})$。该值越高，表明皮损区域被错误分割为正常皮肤的程度越低。

2) 特异度(specificity)又称为真阴性率，定义为：$I_{\mathrm{Specificity}}=V_{\mathrm{TN}}/(V_{\mathrm{TN}}+V_{\mathrm{FP}})$。
该值越高，表明正常皮肤被错误分割为皮损区域的程度越低。

3) Jaccard 指数又称为 Jaccard 相似性系数，定义为：$I_{\mathrm{Jaccard}}=V_{\mathrm{TP}}/(V_{\mathrm{TP}}+V_{\mathrm{FP}}+V_{\mathrm{FN}})$。Jaccard 指数用来衡量样本集之间的相似度，此值越大说明两个样本集之间相似度越高。

4) Dice 指数定义为：$I_{\mathrm{Dice}}=2V_{\mathrm{TP}}/(2V_{\mathrm{TP}}+V_{\mathrm{FN}}+V_{\mathrm{FP}})$。Dice 指数也是用来衡量样本集之间的相似度，与 Jarccard 指数类似。

## 3　实验

### 3.1　数据集

医学图像数据集的标注需要专业医生的参与才能使标签更可靠，然而标注过程费时费力，精细的标注成本

极高,因此本文采用的数据集是国际皮肤成像协会(international skin imaging collaboration,ISIC)于2017年以及2018年举办的挑战赛所使用的比赛数据集ISIC2017和ISIC2018。其中ISIC2017中训练集有2 000张图像,测试集有600张图像,由于ISIC2018数据测试集的标签图片未公布,所以本文采用手动划分的方式,对于ISIC2018中训练集有2 075张皮肤镜图像,测试集有519张图像。

### 3.2 数据预处理

在将图像作为CNN模型的输入之前,使用大小调整、缩放、去除毛发伪影对图像进行预处理。

1)压缩图像尺寸。在将图像输入CNN之前调整图像的大小是一种很好的做法。它可以减少显存的占用,从而消除计算能力限制。皮肤镜图像大小不同,例如ISIC2017数据集就包括540×722、2 000×2 565、4 499×6 748等图像大小,为了克服图像大小的差异,将所有图像及相应的标签下采样至512×512分辨率后再进行处理。所有RBG图像均为JPEG文件格式,而相应的标签为png格式及npy格式。

2)数据标准化。在训练前对图像进行标准化,以消除对比度差的问题。标准化更改像素值的范围,在0和1之间重新缩放图像,以便输入数据在所有维度上都以0为中心。归一化是通过从图像的平均值中减去图像,然后除以图像的标准偏差来获得的。

3)去除毛发伪影结构。皮肤镜图像包含毛发状伪影,在分割病变区域时会带来不便。对图像进行一系列形态学操作以去除这些毛发状结构,然后应用修复算法[22]用相邻像素替换像素值。首先将RGB转换为灰度图像;然后对灰度图像应用黑顶过滤器(black hatfilter)[23];再在生成的二值掩模上实现修复算法;最后用相邻像素修复头发占据的区域。去除毛发过程的相应图像如封三彩图10所示。其中黑顶过滤器是通过从原始图像中减去图像的关闭来获得的,即如果A是原始输入图像,B是输入图像的结束,则黑顶帽过滤器由 $f_{\mathrm{BlackHat}}(A)=(A\times B)-A$ 定义。

闭合形态学操作是对集合A和B的膨胀的侵蚀。闭合填充区域中的小孔,同时保持初始区域大小不变。它保留了与结构元素类似的背景像素,同时消除了背景的所有其他区域。

### 3.3 实验训练过程中细节和参数设置

本次实验采用的是Pytorch 1.10.0框架以及Python 3.7.11。在训练时将训练数据标签转换成png以及npy数据格式进行训练。开始训练时学习率设置成为1e-4,学习率调整策略采用余弦退火调整策略,使用Adam算法作为优化算法,训练过程采取50个Epoch进行训练,Batchsize设为4,patience设定为10,即损失经过10次迭代无下降则停止。

### 3.4 实验对比

为了验证该算法的性能,选取了几个模型:其中包括UNet[2]、UNet++[24]、DeeplabV3[25]、GCN[26]、CeNet[27]以及FCN[28]进行比较。表1展示了模型在ISIC2018数据集上的对比结果。

由表1可知,在ISIC2018图像分割的任务中,本文模型的Jarcard相关系数($I_{\mathrm{Jaccard}}$)比DeeplabV3要高出大约0.4个百分点,Dice相关系数($I_{\mathrm{Dice}}$)高出大约0.2个百分点,是这几种模型之中最高的。

表2展示了模型在ISIC2017数据集上的对比结果。由表2可知,在ISIC2017图像分割的任务中,Proposed模型的Jarcard相关系数($I_{\mathrm{Jaccard}}$)比DeeplabV3高出大约0.1个百分点,Dice相关系数也高出大约0.1个百分点,是这几种模型之中最高的。

**表1 ISIC2018数据集上的对比实验结果**

**Tab.1 Experimental results on ISIC2018**

| 模型 | $I_{\mathrm{Jaccard}}$ | $I_{\mathrm{Dice}}$ | $I_{\mathrm{Sensitive}}$ | $I_{\mathrm{Specificity}}$ |
|---|---|---|---|---|
| 本文模型 | 0.843 689 | 0.908 548 | 0.915 274 | 0.978 615 |
| DeeplabV3 | 0.800 653 | 0.882 353 | 0.913 341 | 0.964 502 |
| GCN | 0.787 655 | 0.869 157 | 0.914 693 | 0.957 608 |
| UNet | 0.765 696 | 0.837 971 | 0.888 161 | 0.957 158 |
| UNet++ | 0.816 523 | 0.879 135 | 0.907 981 | 0.965 687 |
| CENet | 0.802 139 | 0.882 381 | 0.897 632 | 0.961 263 |
| FCN | 0.755 641 | 0.836 683 | 0.842 002 | 0.951 485 |

**表2 ISIC2017对比实验结果**

**Tab.2 Experimental results on ISIC2017**

| 模型 | $I_{\mathrm{Jaccard}}$ | $I_{\mathrm{Dice}}$ | $I_{\mathrm{Sensitive}}$ | $I_{\mathrm{Specificity}}$ |
|---|---|---|---|---|
| 本文模型 | 0.767 901 | 0.854 534 | 0.921 502 | 0.963 473 |
| DeeplabV3 | 0.763 729 | 0.846 915 | 0.853 650 | 0.959 699 |
| GCN | 0.757 065 | 0.843 708 | 0.885 517 | 0.940 439 |
| UNet | 0.680 358 | 0.767 521 | 0.825 564 | 0.949 155 |
| UNet++ | 0.742 200 | 0.830 400 | 0.839 200 | 0.978 400 |
| CENet | 0.760 000 | 0.844 600 | 0.823 500 | 0.975 200 |
| FCN | 0.710 868 | 0.796 294 | 0.843 876 | 0.946 012 |

封三彩图 11 展示了分割结果，由于对比算法较多，所以图中只展示了本文算法与 Unet 算法的分割效果，从首尾两列的对比结果来看，可以看出本文所提的算法更加关注部分细节轮廓。

## 4　结论

在本研究中，将 CNN 和 Transformer 结构相结合，提出了一种新的皮肤镜图像分割算法。首先对皮肤病变图像进行毛发去除预处理，减少毛发噪声对结果的影响，然后采用 Resnet 作为主干特征提取网络提取特征，将提取到的特征图序列作为 Transformer 的输入，在 Transformer 中加入了新的结构边界注意门以提取足够的局部细节来处理模糊边界，最后采用 DenseASPP 增强特征表示和处理多尺度信息。采用一种考虑一种新的损失函数以增强对边界关注，从而解决了病变区域形状不规则以及边界模糊问题，通过与不同的模型之间对比以及效果展示证明了本文所提算法有效性。

**参考文献：**


[1] LONG J，SHELHAMER E，DARRELL T. Fully convolutional networks for semantic segmentation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston：IEEE，2015：15524435.

[2] RONNEBERGER O，FISCHER P，BROX T. U-Net：convolutional networks for biomedical image segmentation[C]//Internet Conference on Medical Image Computing and Computer-Assisted Intervention. Cham：Springer：2015：234-241.

[3] WEN H D. II-FCN for skin lesion analysis towards melanoma detection[EB/OL]. (2017-03-14)[2022-04-11]. https：//arxiv. org/abs/1702. 08699.

[4] SZEGEDY C，VANHOUCKE V，IOFFE S，et al. Rethinking the inception architecture for computer vision[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas：IEEE，2016：2818-2826.

[5] SUTTON C，MCCALLUM A. An introduction to conditional random fields[J]. Foundations and Trends in Machine Learning，2012，4(4)：267-373.

[6] THAO L T，QUANG N H. Automatic skin lesion analysis towards melanoma detection[C]//2017 21st Asia Pacific Symposium on Intelligent and Evolutionary Systems. Hanoi：IEEE，2017：17452688.

[7] YU L Q，CHEN H，DOU Q，et al. Automated melanoma recognition in dermoscopy images via very deep residual networks[J]. IEEE Transactions on Medical Imaging，2017，36(4)：994-1004.

[8] HE K M，ZHANG X Y，REN S Q，et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas：IEEE，2016：770-778.

[9] CHEN J N，LU Y Y，YU Q H，et al. TransUNet：transformers make strong encoders for medical image segmentation[EB/OL]. (2021-02-08)[2022-04-11]. https：//arxiv. org/abs/2102. 04306.

[10] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[EB/OL]. (2017-12-06)[2022-04-11]. https：//arxiv. org/abs/1706. 03762.

[11] YANG M K，YU K，ZHANG C，et al. DenseASPP for semantic segmentation in street scenes[C]//2018IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City：IEEE，2018：18347697.

[12] KAUR R，GHOLAMHOSSEINI H，SINHA R，et al. Automatic lesion segmentation using atrous convolutional deep neural networks in dermoscopic skin cancer images[J]. BMC Medical Imaging，2022，22(1)：1-13.

[13] BOGO F，PERUCH F，FORTINA A B，et al. Where's the lesion? Variability in human and automated segmentation of dermoscopy images of melanocytic skin lesions[M]//CELEBI M E，MENDONCA T，MARQUES J S. Dermoscopy image Analysis. Montrouge：CRC Press，2015：67-96.

[14] HARDIE R C，ALI R，De SILVA M S，et al. Skin lesion segmentation and classification for ISIC 2018 using traditional classifiers with hand-crafted features[EB/OL]. (2018-07-18)[2022-04-11]. https：//arxiv. org/abs/1807. 07001.

[15] 齐永锋，侯璐璐，段友放. 基于 DenseNet-BC 网络的皮肤镜下皮肤损伤分割[J]. 计算机工程与科学，2020，42(6)：1060-1067.
QI Y F，HOU L L，DUAN Y F. Segmentation of skin lesions under dermoscopic skin damage based on the SenseNet-BC network[J]. Computer Engineering and Science，2020，42(6)：1060-1067.

[16] SIMONYAN K，ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2022-04-11]. https：//arxiv. org/abs/1409. 1556.

[17] SZEGEDY C，LIU W，JIA Y，et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston：IEEE，2015：15523970.

[18] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[EB/OL]. (2016-08-23)[2022-04-11]. https://arxiv.org/abs/1602.07261.

[19] FU J, LIU J, TIAN H J, et al. Dual attention network for scene segmentation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 3146-3154.

[20] HE K M, ZHANG X Y, REN S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[21] 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, 2018, 40(4): 834-848.

[22] TELEA A. An image inpainting technique based on the fast marching method[J]. Journal of Graphics Tools, 2004, 9(1): 23-34.

[23] THAPAR S, GARG S. Study and implementation of various morphology based image contrast enhancement techniques[EB/OL]. (2012-01-31)[2022-04-11]. https://www.researchmanuscripts.com/isociety2012/42.pdf.

[24] ZHOU Z W, SIDDIQUEE M M R, TAJBAKHSH N, et al. UNet++: redesigning skip connections to exploit multiscale features in image segmentation[J]. IEEE Transactions on Medical Imaging, 2020, 39(6): 1856-1867.

[25] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. (2017-12-05)[2022-04-11]. https://arxiv.org/abs/1706.05587.

[26] PENG C, ZHANG X Y, YU G, et al. Large kernel matters-improve semantic segmentation by global convolutional network [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 17355589.

[27] 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.

[28] VILLA M, DARDENNE G, NASAN M, et al. FCN-based approach for the automatic segmentation of bone surfaces in ultrasound images[J]. International Journal of Computer Assisted Radiology and Surgery, 2018, 13(11): 1707-1716.

## Application of CNN Transformer Network in Dermoscopy Image Segmentation

DONG Yumin, WEI Lixing

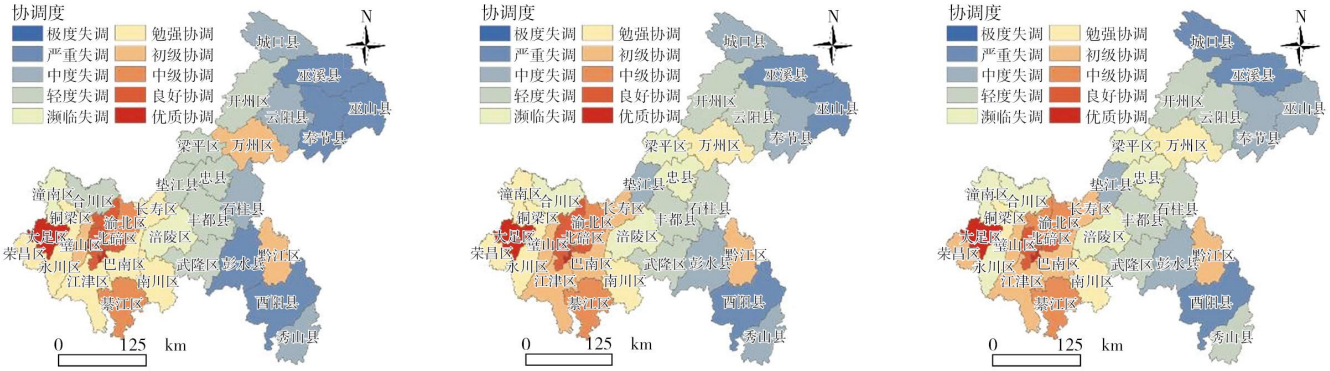
(College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China)

**Abstract**: [**Purposes**]Aiming at the irregular shape, blurred boundaries and hair artifacts of skin lesions in skin lesion images, it proposes a skin lesion segmentation algorithm combining CNN and Transformer. [**Methods**]Firstly, the skin lesion image was pretreated for hair removal to reduce the influence of hair noise on the result, and then a segmentation model combining CNN and Transformer was constructed, using Resnet as the backbone feature extraction network to extract features, and the extracted feature map sequence was used as the input of Transformer, and a new structural boundary attention gate was added to the Transformer to extract enough local details to process the blurry boundary. Finally, The DeanseASPP enhanced feature is used to represent and process multi-scale information, and an improved loss function is proposed, the purpose of which is to make the model focus on the boundary region part when calculating the loss function. [**Findings**]The experimental results show that the dice value and JI value are 0.854 534 and 0.767 901 on the ISIC2017 dataset, and 0.908 548 and 0.843 689 on the ISIC2018 dataset, respectively, which achieves good results compared with other advanced models. [**Conclusions**]Its effectiveness is proved by comparing with different models and showing the effect.

**Keywords**: CNN; Transformer; DenseASPP; segmentation of skin lesions

(责任编辑 许 甲)

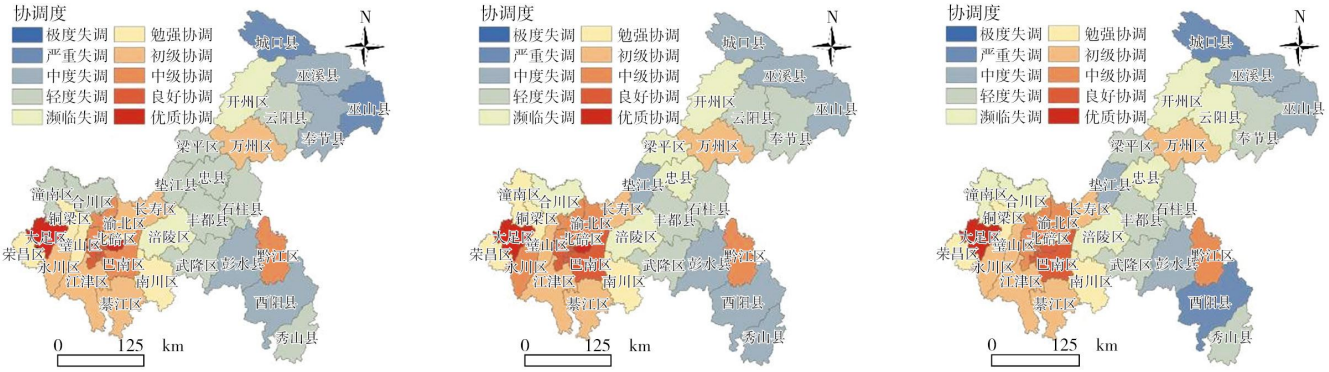
（接正文 79 页）




a 2015 年　　b 2017 年　　c 2019 年

图 1　研究区经济 – 城镇化协调度

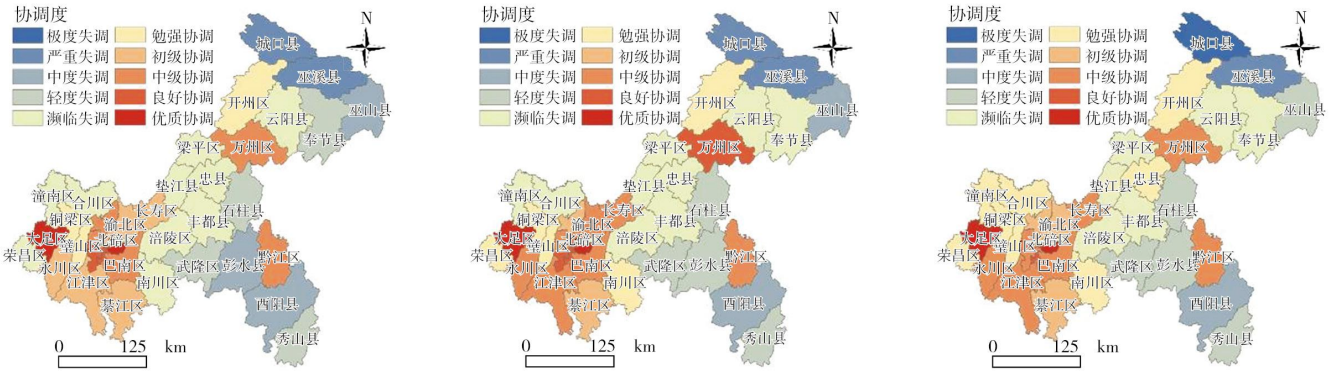
**Fig.1　The coordination degree between and urbanization in the study area**




a 2015 年　　b 2017 年　　c 2019 年

图 2　研究区城镇化 – 碳排放协调度

**Fig.2　The coordination degree between urbanization and carbon emission in the study area**

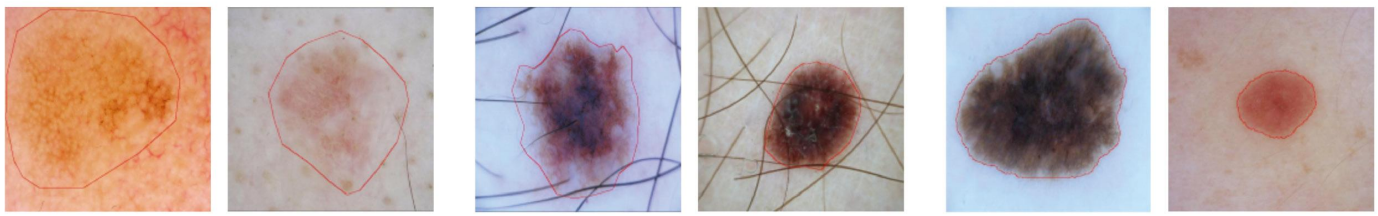



a 2015 年　　b 2017 年　　c 2019 年

图 3　研究区经济 – 碳排放协调度

**Fig.3　The coordination degree between economy and carbon emissions in the study area**

（接正文127页）



a 病变皮肤与周围对比度较低　　b 图像中存在纹理和毛发等燥声　　c 病变区域不规则且边界模糊

图 1　病变区域存在的问题

**Fig.1　Problems in the lesion area**

（接正文127页）

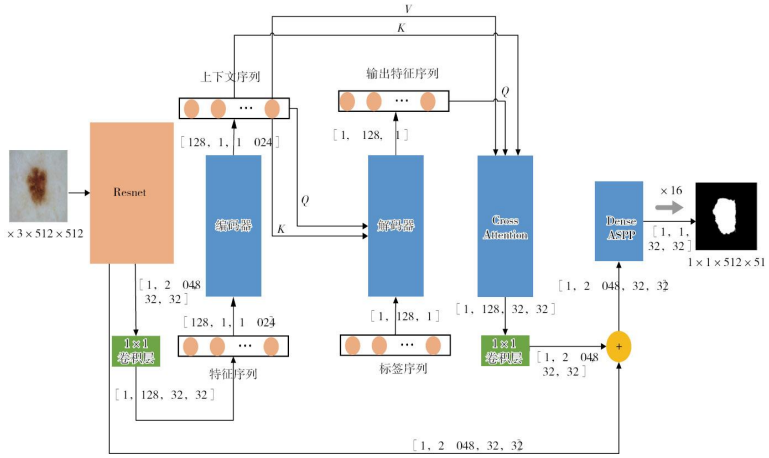



图 2　模型整体结构图

**Fig.2　Overall structure diagram of the model**

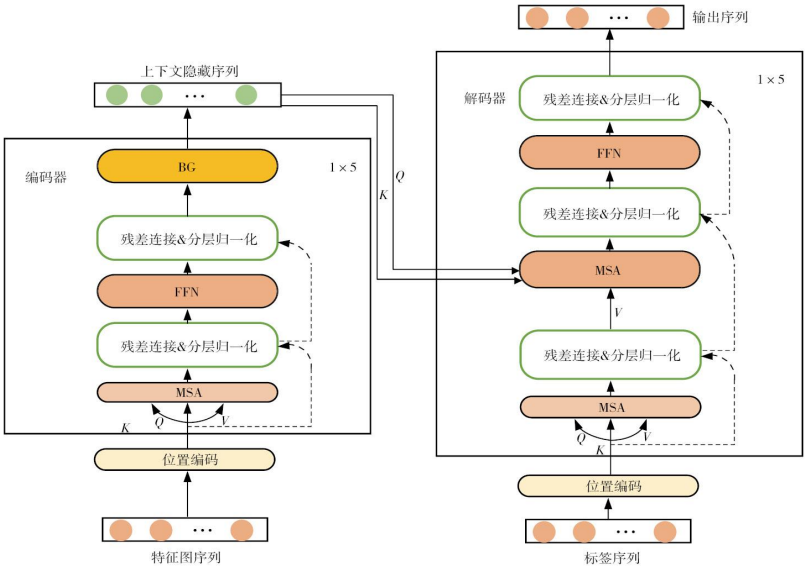
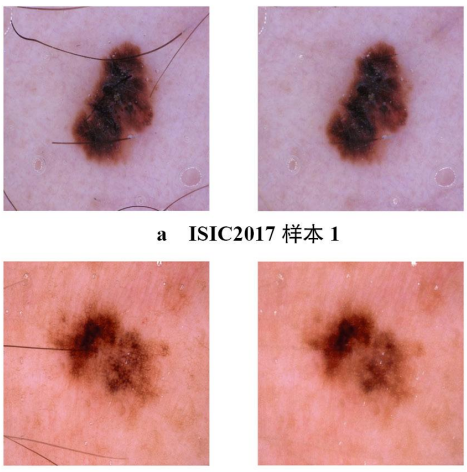



图 3　Transformer 总体架构图

**Fig. 3　Overall architecture diagram of Transformer**

（接正文132页）



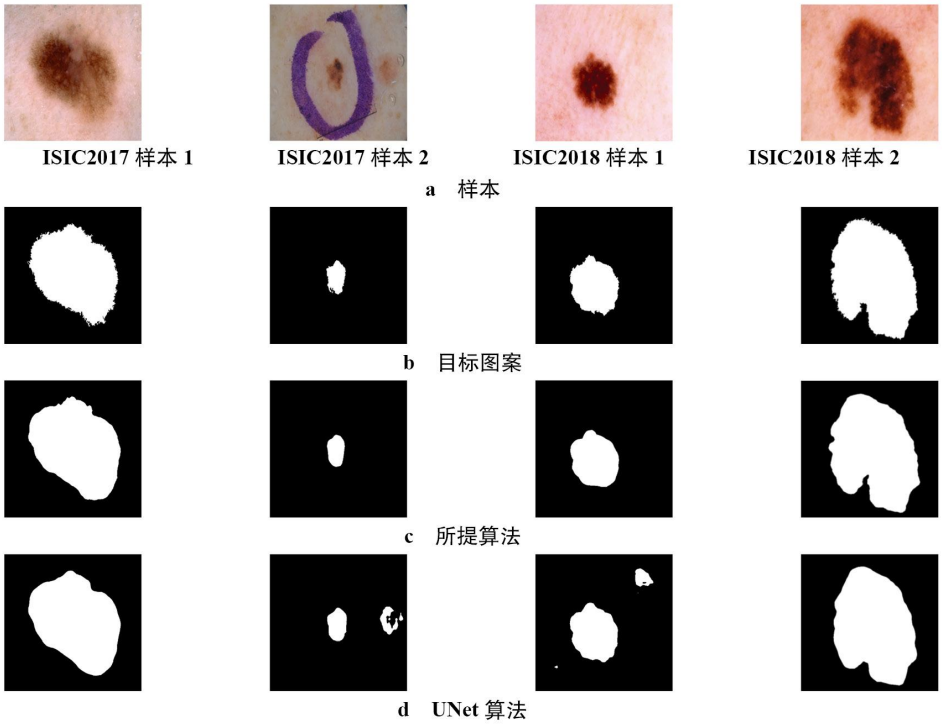
a　ISIC2017 样本 1

b　ISIC2018 样本 2

图 10　去除毛发过程的相应图像

**Fig. 10　Corresponding image of the hair removal process**

（接正文133页）



ISIC2017 样本 1　ISIC2017 样本 2　ISIC2018 样本 1　ISIC2018 样本 2

a　样本

b　目标图案

c　所提算法

d　UNet 算法

图 11　实验对比效果图

**Fig. 11　Experimental comparison effect diagram**