深度学习技术在医学图像分析应用中的发展

王韧立[1] 王茹奕[1]

1 南方科技大学 广东省 深圳市 518055
（12112321@mail.sustech.edu.cn）

摘 要 医学图像分析，是指应用医学影像，人工智能，数值算法等技术，对医学图像提取关键信息等操作。影像技术在医学的诊断与治疗中正发挥着日益重要的作用。当前，基于深度学习的算法被广泛地运用于医学图像分析中。文章通过总结研究相关的文献资料，对医学图像分割与分析技术的发展过程进行了大致梳理，并通过对比分析卷积神经网络和 transformer 这两种主要图像分析方法，揭示了二者的适用场景 、优势与不足，同时探究了其近年来的发展趋势，预测了其未来可能的研究方向，对深度学习技术在医学图像分析应用中的发展进行了较为全面的描绘与总结。

关键词： 医学图像分割；深度学习；卷积神经网络；Transformer；图像处理

R445

Development of Deep Learning in Medical Image Analysis

WANG Ren-li[1]，WANG Ru-yi[1]

1 Southern University of Science and Technology， Shenzhen 518055，China

Abstract Medical image analysis refers to the application of medical imaging, artificial intelligence, numerical algorithms and other technologies to extract key information from medical images. Imaging technology is playing an increasingly important role in the diagnosis and treatment of medicine. At present, algorithms based on deep learning are widely used in medical image analysis. By summarizing the relevant literature of the research, this paper roughly sorts out the development process of medical image segmentation and analysis technology, and reveals the applicable scenarios, advantages and disadvantages of convolutional neural network and transformer by comparing and analyzing the two main image analysis methods, and explores its development trend in recent years, predicts its possible future research direction, and comprehensively describes and summarizes the development of deep learning technology in medical image analysis applications.

Keywords Medical image segmentation, Deep learning, Convolutional neural networks, Transformer, Image processing

1 引言

医学图像分析，是指应用医学影像，数字建模，人工智能，数值算法等学科交叉领域的技术，对医学图像进行分割，提取关键信息等操作。影像技术已经日益成为医疗机构疾病诊断手术计划和预后评估的重要手段之一[1]。精确分析图像的关键区域可以帮助医生对疾病作出正确的判断。在临床决策中，图像分割可为计算机辅助诊断和治疗提供可靠的依据，对于定量分析和手术指导也具有重要作用。与此同时快速和准确的图像分割，也是三维可视化、定量分析等后续环节开展之前需要进行的重要步骤，为图像引导手术、治疗评估和放疗计划等重要临床应用奠定了最根本的基础[2,3]。近年来，基于深度学习的自动分割算法取得了较为迅速的进展。在早期的医学图像分割研究中，采用的主要是直接进行

到稿日期： 返修日期：
基金项目：XX 基金（基金号）；YY 项目（项目号）
This work was supported by XX (No.) and YY (No.).
通信作者：姓名（E-mail）

基金项目名称、编号的英文翻译确保无误。

图像处理的一些经典方法，例如基于灰度的区域增长算法以及边界提取等，由于图像包含有众多不同的信息，传统的方法容易遇到计算缓慢，精度不高等的问题，随着技术手段的发展，计算机视觉相关理论的丰富，诸如卷积技术，transformer 以及一些融合了多种技术的先进方法都被运用于医学图像处理之中[4]。本文将大致梳理医学图像分割技术的发展过程，并对不同的图像处理技术进行对比分析，探究其近年来以及未来的发展方向，以期对医学图像的分析形成较为完整准确的认识。

2 深度学习在医学图像分析中的发展

2.1 传统图像处理方法

在传统的医学图像分析领域，从感兴趣区域（ROI）提取特征并应用图像分类算法进行图像的识别与分类是一个重要方法[5]。图像分类算法，顾名思义，是一种根据输入图像描述的内容进行图像分类的方法。它有五个基本步骤，即图像数据集输入、图像预处理、特征提取、分类器训练和图像识别。

2.1.1 图像数据集的输入

首先，将两类数据集，训练集和测试集，输入电脑。训练集中的图像已经被分类和标记，是供计算机学习数据特征的；而测试集则包含了大量各种类型的图像，用来测试计算机是否已经训练完毕。

2.1.2 图像预处理

在输入图像数据集后，需要先对图像进行预处理，提高图像质量，减少一些不必要的干扰，使计算机能够更好地学习特征。由于计算机没有视觉，不能直接识别图像，因此有必要研究如何通过各种处理来提取有用的信息，以获得图像的"非图像"表示，如数值、向量和符号等。然后，我们就可以通过训练过程教会计算机如何理解这些特征，从而使计算机具备识别图像的能力。

2.1.3 特征提取

在提取时，通常要根据各类图像的特点选择最佳的特征提取算法，下面就介绍两种典型的算法。

2.1.3.1 LBP 特征提取

LBP 利用了结构化思想提取图像分割的各个区域的特征，然后对各个区域进行统计，作为提取的最终图像特征。 早期的 LBP 算子定义在一个 3*3 的区域内，以这个区域的中心像素为标准值，将其与相邻的 8 个区域的灰度值进行比较，得到一个 8 位的二进制数，即得到该区域的 LBP 值，这个值也反映出这个区域的纹理信息。

方法如下：首先，将待提取的图像划分为 16*16 个区域。然后，对于每个区域中的一个像素，通过上面提到的获取 LBP 值的方法获得该区域中心像素的 LBP 值。随后，根据该值计算出每个部分的直方图，即每个数字的出现次数，然后对直方图进行归一化处理。最后，将各部分得到的直方图衔接成一个特征向量，也就是整个图像的 LBP 纹理特征向量。

2.1.3.2 HOG 特征提取

HOG 是计算机视觉和图像处理中用于物体检测的一种特征描述符。它通过计算和统计图像局部区域的梯度方向直方图构成特征。它的主要思想是，图像中局部目标的外观和形状可以由梯度或边缘的方向密度分布很好地描述。步骤如下：

第一， 对一幅图像进行灰度化处理。

第二，用伽马校正法对输入图像的颜色空间进行归一化（normalize），目的是调整图像的对比度，减少图像中局部阴影和光照变化造成的影响，同时可以抑制噪声的干扰。

第三，计算图像每个像素的梯度（包括大小和方向），主要目的是捕捉轮廓信息，同时进一步弱化光照的干扰。

第四，将图像划分为小单元。

第五，统计每个单元的梯度直方图（不同梯度的数量），形成每个单元的描述符。

然后，每隔几个单元形成一个区块，将一个区块中所有单元的特征描述符串联起来，得到该区块的 HOG 特征描述符。

最后，图像的 HOG 特征描述符可以通过连接图像中所有块的 HOG 特征描述符得到。这就是最终的特征向量，可以用于分类。

2.1.4 分类器训练

首先，使用训练集中的图片数据，知道训练集中每个图片数据的分类标签。然后，根据这些已知前提，找到相应的判断函数或标准，设计判断函数模型。最后，根据训练集中的数据来确定函数模型中的参数。

2.1.5 图像识别

分类器被训练完毕后，就可以对要分类的图像进行分类和识别了。

2.2 以卷积神经网络（Convolutional Neural Networks, CNN）为基础的医学图像分析方法

2.2.1 卷积神经网络的介绍，定义及其由来

卷积神经网络是用于分析视觉图像的一种常用的深度学习模型，是深度学习的代表算法之一，通常指代具有类似于人工神经网络系统结构的一种多层感知器，包含卷积计算且具有深度结构。卷积神经网络之灵感主要源于神经科学中对于人类视觉系统中的视觉皮层的研究。在视觉皮层中，感光细胞的种类不同，其对于特定区间的敏感程度也存在差异，卷积神经网络借鉴并发展了这一特点以更好地处理图像[6]。于 1988 年，Wei Zhang 提出了第一个用于医学图像分析检测的二维卷积神经网络，为平移不变人工神经网络（SIANN）。在深度学习的相关理论于 2006 年被提出之后，卷积神经网络的表征学习能力得到了较为广泛的关注[7]。

2.2.2 卷积神经网络的基本结构

一般而言，卷积神经网络的主要结构由数据输入层，卷积计算层，池化层以及全连接层这四个层级构成。其中卷积层可以进一步细分出卷积核，卷积层参数，激励函数等要素。在部分特殊的卷积神经网络中，各层级的顺序和包含关系可能会有所变化。原始输入的图像信息经过各层级的处理，提取出关键的特征。

2.2.3 卷积神经网络的结构分析和运算步骤

卷积神经网络的输入层是能够处理多种维度的数据的，不过，在医学图像分析相关的卷积运算中，常见的输入数据类型为二维或三维，在该层的主要工作是对原始的二维输入图像进行简单的预处理，对于 nxn 的二维图像数据，输入依旧为 nxn 的二维神经元，而对于 RGB 格式的图像数据，输入则为 3xnxn 的三维神经元。在卷积层中，主要进行的工作是对输入层的数据进行特征提取，以二维为例，用卷积核按照既定的次序规律扫过输入数据构成的矩阵区域（若对边缘数据进行滤波，则可对输入值边缘进行零填充后泛卷积），在卷积核感受野内的输入数值，与卷积核上的权重进行矩阵元素的乘法并求和，再将得到的加权和叠加上偏差值，若数据线性不可分，则可通过激励函数的变换变为可分（如求平方值等）。在池化层，主要的工作是对上一步中得到的过多数据进行压缩和简化，常见的方法有将数据划分成若干个子区域后，对各子区域取平均值（平均池化）或最大值（最大池化）。在全连接层，主要的工作是对池化层的数据按照特征进行整合分类，可以通过全连接将卷积后得到的二维特征图转化为一维的向量。总体而言，卷积神经网络各步骤间环环相扣，从而对原始数据进行了简化和特征的提取[8]。一些不同的用于医学图像分割的卷积神经网络被广泛地提出，如 AlexNet[9]，GoogleNet[10]，SequeezeNet[11]等，这些卷积模型主要的区别在于引入了不同数目的卷积和池化层数，并运用了相应的运算方式，从而产生了不尽相同的图像分割效果[12]。

2.2.4 卷积神经网络在医学图像分割与分析中存在的不足和改进

首先卷积神经网络中的全连接层无法处理不同大小的输入数据，具有全连接层的卷积神经网络并不能被很好地用于具有不同尺寸的图像分割的任务，因为在该任务中，关键要素出现的数目是不固定的，因此，输出层的长度不能是定值。一种对上述缺陷的改进方案是采用全卷积神经网络（Fully Convolutional Network,FCN）[13]。全卷积网络与传统的卷积神经网络主要的区别在于前者中所有的层都为卷积层，用卷积网络替换了传统的全连接网络，因此，取消了全连接层对输入的神经元个数的限制，从而使得卷积层的输入可以是不同尺寸的图像。在全卷积操作的最后过程中，可以通过转置卷积层进行上采样将特征图的尺寸扩张为原始输入图像的尺寸大小，解决了卷积和池化导致尺寸变小的问题[12]。然而，传统的全卷积仍然存在一定的局限性，例如实际速度仍然较慢以及难以考虑全局信息的关联。同时，由于每个输出值对应的感受野是固定的，导致图像分割效果不足。在一些研究中提出了较为完善的全卷积模型，例如ParseNet[13]，采用了全局均值池化等方法，从而在一定程度上考虑了全局信息的关联[12]。

2.3 以 transformer 为基础的图像处理方法

2.3.1 定义

Ashish Vaswani 在 2017 年的《Attention Is All You Need》中提出相比于主流的基于复杂的循环、包含 encoder 和 decoder 的卷积神经网络，通常在 encoder 和 decoder 之间使用注意力机制（Attention）的模型性能更优，于是他彻底抛弃 RNN，提出了完全基于注意力机制的 transformer 算法[14]。

如图 1 所示，在以往的经典模型中一般使用两个 RNN，一个作为编码器，一个作为解码器。编码器的作用是将输入数据编码成一个特征向量，解码器则将这个特征向量解码成预测结果。

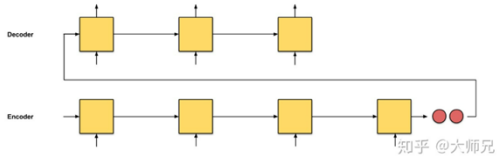

图 1

Fig. 1

这个模型的问题是只将编码器的最后一个节点的结果进行了输出，但是对于一个序列长度特别长的特征来说，这种方式无疑将会遗忘大量的前面时间片的特征。既然如此，我们不如将每个时间片的输出都提供给解码器，给解码器提供更好的特征。那么解码器如何使用这些特征就是我们这里介绍的 Attention 的作用，如图 2 所示

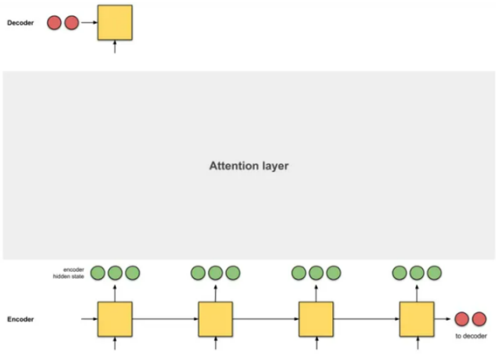

图 2

Fig. 2

在这里，Attention 是一个介于编码器和解码器之间的一个接口，用于将编码器的编码结果以一种更有效的方式传递给解码器。一个特别简单且有效的方式就是让解码器知道哪些特征重要，哪些特征不重要，即让解码器明白如何进行当前时间片的预测结果和输入编码的对齐，如图 3 所示。Attention 模型学习了编码器和解码器的对齐方式，因此也被叫做对齐模型（Alignment Model）。

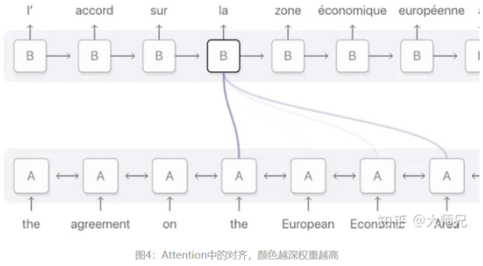

图4：Attention中的对齐，颜色越深权重越高

图 3

Fig. 3

而在 2020 年 Alexey Dosovitskiy 提出了将 Transformer 直接应用于图像领域，进行图像分类任务，而不修改 Transformer 架构，也不使用 CNN 的思路。其主要想法是将 image 将图像分割成大小一致的 patch 小块，并将这些块转化为线性嵌入序列，作为 Transformer 的输入。在 patch 前加入一个 class 表示，对应全局信息（图片分类），进行最终的预测，如图 4 所示[15]。

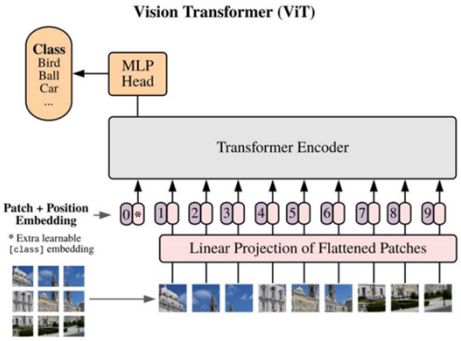

图 4

Fig. 4

2.3.2 attention 算法步骤[16]

步骤一，生成编码节点。将输入数据依次输入到 RNN 中，得到编码器每个时间片的隐层状态的编码结果，并将编码器的最后一个输出作为解码器的第一个状态的输入（decoder hidden state）。

步骤二，为每个编码器的隐层状态计算一个得分。使用当前编码器的当前时间片的隐层状态和解码器的隐层状态计算一个得分，得分的计算方式有多种，如点乘操作等。

步骤三：使用 softmax 对得分进行归一化。将 softmax 作用到步骤 2 得到的得分上，得到和为 1 的分数。

步骤四：使用分数对隐层状态进行加权。将上述的分数与隐层状态进行点乘操作，得到加权之后的特征，这个特征也叫做对齐特征（Alignment Vector）或者注意力特征（Attention Vector）。

步骤五：对特征向量进行求和。这一步是将加权之后的特征进行加和，得到最终的编码器的特征向量。

步骤六：将特征向量应用的解码器。最后一步是将含有 Attention 的编码器编码的结果提供给解码器进行解码，注意每个时间片的 Attention 的结果会随着 decoder hidden state 的改变而更改。

3 对于医学图像分割中各个方法的比较分析

3.1 对比对象

卷积神经网络以及 transformer 都是医学图像分析中得到广泛运用的模型，两者有许多的共同之处，但在算法和效果上存在明显的差异性，接下来将着重对这两者的异同点进行比较分析。

3.2 产生不同的方面和内容

3.2.1 参数和数据规模

卷积神经网络所学习的参数主要是静态参数在经过一定数量的学习后即可保持不变，而

transformer 所具有的参数是动态的，可以随着数据量的增加而不断调整。随着计算机视觉相关技术的发展，需要处理的数据规模不断扩大，一般而言，大多数的卷积神经网络模型更适用于监督性学习，在面对小规模数据时即可发挥作用，不过在面对海量数据时，其适配能力并不突出，与此同时，根据 Dosovitskiy 等人的研究可知，transformer 由于需要对空间关系进行预训练学习，在面对大数据时具有更好的适配能力，经过适当的优化，随着数据量的增加，其模型的表现更好，优势更加突出[17]。不过，transformer 无法直接利用原始图像尺度，平移不变性，等先验知识，依赖后期的学习，在数据集较小的情况下，适用性弱于卷积神经网络。

3.2.2 关键点和自注意力机制

对于卷积神经网络而言，由于卷积核上的权重是被输入值共享的，因此卷积能够很好地捕获相邻区域的特征，且具有平移不变性。同时，在卷积的处理过程中，相邻的数据点之间被更加紧密地关联在了一起，从而具有高度的局部性。而 transformer 具有的多头自注意力机制，使其能够高效地处理长程依赖关系，更多地学习到了所有特征之间的相互关系，不完全局限于分散的数据本身，而是将局部的信息推广到了全局的信息，具备更加优越的普适性[18,19]。如图 5 所示为一项研究中分别运用卷积神经网络模型和 transformer 模型分别对于肝脏血管的图像进行分析运算后的效果[18]。通过图片可以注意到，由于卷积神经网络的局部性，难以关联全局信息，部分原始图像中连结为一个整体的部分被错误地断开了，而以 Visual Transformer 为基础的输出图像，这些部分仍然保持连结，并没有出现 CNN 中出现的异常断开的问题，正确地反映了全局依赖的关系。

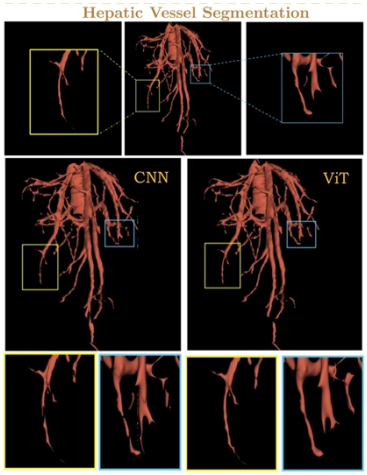

图 5

Fig. 5

3.2.3 层级之间的相似程度

Maithra Raghu 等研究者对卷积神经网络和 Visual Transformer 中的一些关键层级的 Centered Kernel Alignment (CKA) 进行了检验，用以大致研究在随机初始化和不同程度的训练情况下，两种神经网络模型各自的各个层级间的关联程度[20]。该研究表明，在底层（接近输入的层级）中，两者表征的内容十分相似，然而，由于两个模型基于表征截然不同的处理方式，经过层层的运算，最终上层（接近输出的层级）中的表征具有相当大的差异[20]。同时，研究人员还发现，Visual Transformer 中，数据从底层到达高层的相似程度要普遍高于卷积神经网络[20]。经过其分析，产生这种现象的原因主要在于 Visual Transformer 中的跳跃连接结构较为成功地保护了表征信息从底层到高层的传递，实验表明，若撤去特定步骤上的跳跃连接结构，相应的表征信息就无法得到传递[20]。据此，该项研究的结果表明，对于更高层级的表征而言，transformer 相对于常见的卷积神经网络（ResNet）能够更为细致地保留局部空间位置信息。实验中 Vit 与 ResNet 的结果的对比图如下

图所示[20]。

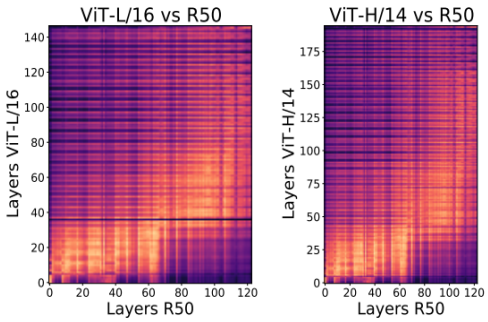

图 6

Fig. 6

4 未来发展方向

随着人们对于 Transformer 的研究不断深入，其优势也不断展现，不断有研究者发现 transformer 在性能与理解自然语言等方面的优势。这不经令人疑惑：在未来 transformer 是否能完全取代 CNN。

其实，Transformer 和 CNN 二者可能并非取代和被取代的关系，而是互相融合，取长补短。

目前 Transformer 仍有许多问题需要解决，首先，transformer 模型的参数量一般比较大，在训练和推理的计算成本较高，在 CV 中还需要更多的探索；其次，transformer 需要使用大量的训练数据进行预训练，而在某些领域，特别是本文关注的医学图像领域，数据量较小，如何高效地利用数据仍是一个需要解决的问题；然后，transformer 对输入图像的处理不太符合直觉，其理论原因也还不太清楚，因此还需要更多的改进与理论分析[21]。

由于 CNN 和 Transformer 各有优势和不足，二者融合的做法出现在很多 Transformer 的跨界论文中。

CNN 网络在提取底层特征和视觉结构方面有比较大的优势。这些底层特征构成了在 patch 上的关键点、线和一些基本的图像结构。这些底层特征具有明显的几何特性，往往关注诸如平移、旋转等变换下的一致性或者说是共变性。CNN 网络在处理这类共变性时是很自然的选择。但当我们检测得到这些基本视觉要素后，高层的视觉语义信息往往更关注这些要素之间如何关联在一起进而构成一个物体，以及物体与物体之间的空间位置关系如何构成一个场景，这些是我们更加关心的。目前来看，transformer 在处理这些要素之间的关系上更自然也更有效[22]。但是，反过来说，如果全部将 CV 任务中的 CNN 换成 Transformer，我们会遇到很多问题，比如计算量、内存占用量过大等。

总之二者的结合还有很大的探索空间，现有的 Visual Transformer 都还是将原有的 Transformer 结构套到视觉任务做了一些初步探索，未来针对 CV 的特性设计更适配视觉特性的 Transformer 将会带来更好的性能提升；而且现有的 Visual Transformer 一般是一个模型做单个任务，近来有一些模型可以单模型做多任务，未来可以尝试有一个世界模型，处理所有任务。

5 总结

医学图像分析是一个不断发展进步的过程，对医学图像的处理涉及许多学科的交叉领域，相关的算法与模型也在不断地更新与迭代，层出不穷。本文主要介绍了医学图像分析中算法的发展过程，并着重对于卷积神经网络以及以 Visual Transformer 为基础的图像处理方法进行了研究和对比。其中，卷积神经网络借鉴了视觉系统中的视觉皮层，包含卷积计算，具有深度结构，并通过替换全连接层为卷积层的全卷积模型，实现了对任意尺寸的输入图像进行处理，再通过上取样的方式，还原输出为原始输入相同的尺寸，实现了图像的分析。Transformer 由 Encoder 和 Decoder 两个部分组成，利用了由多个自注意力层构成的多头自注意力机制，有效提取和记忆了长距离的信息，减少了计算量并大大提高了并行效率，在医学图像的分析中发挥了重要的作用。

卷积神经网络和 transformer 在图像处理上具有较大的算法区别和各自更适用的情况，由于充分利用了平移不变性和局部特征等先验条件，卷积神经网络在数据集较小的情况下具备优势，效率更高；而由于多头自注意力机制下处理长程依赖关系的能力以及需要对足够数据进行预学习的特点，transformer 在面对更大的数据规模时具有很好的适配性。同时，在多层级的情况下，Visual Transformer 在接近输出端能保留比主流的卷积神经网络更加精细的空间位置信息，从而具有相对更强的性能和实用性。在未来的发展过程中，与 transformer 相关的研究与优化可能会进一步引起广泛关注，同时部分卷积神经网络相关的运算方式也可能被运用到其中，对医学图像的处理方式将会更加丰富和高效。

结束语

本文梳理了医学图像析技术的总体发展过程，着重介绍了以卷积神经网络和 transformer 为基础的模型，并比较分析了二者的异同点和各自的优势，探索了可能的发展趋势。由于知识水平的局限，文章存在着一些不足，例如卷积神经网络和 transformer 主要分析的是基础的模型，一些特殊模型的算法步骤并没有给出详细的介绍，所用于分析特点和对比优缺点的样本数量过小等。未来可能会对不同的具体模型进行更加深入的研究。

参考文献

[1]Budd S, Robinson EC, Kainz B (2021) A survey on active learning

and human-in-the-loop deep learning for medical image analysis.

Med Image Anal 71:102062

[2]Tang, Z., Duan, J., Sun, Y. *et al.* A combined deformable model and medical transformer algorithm for medical image segmentation. *Med Biol Eng Comput* (2022). https://doi.org/10.1007/s11517-022-02702-0

[3]Du G, Cao X, Liang J et al (2020) Medical image segmentation based on U-Net: a review. J Imaging Sci Technol 64:20508

[4]Liu L, Cheng J, Quan Q et al (2020) A survey on U-shaped networks in medical image segmentations. Neurocomputing 409:244–258

[5] Hegde, R. B., Prasad, K., Hebbar, H., & Brij Mohan, K. S. (2019). Feature extraction using traditional image processing and convolutional neural network methods to classify white blood cells: A study. *Australasian Physical & Engineering Sciences in Medicine, 42*(2), 627-638.

[6] P. Malhotra, S. Gupta, and D. Koundal, “Computer aideddiagnosis of pneumonia from chest radiographs,” Journal of Computational and Deoretical Nanoscience, vol. 16, no. 10,pp. 4202–4213, 2019.

[7]G. E. Hinton*, R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks[J] Science 28 Jul 2006: Vol. 313, Issue 5786, pp. 504-507

[8]G. Litjens, T. Kooi, B. E. Bejnordi et al., “A survey on deep learning in medical image analysis,” Medical Image Analysis, vol. 42, pp. 60–88, 2017.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Proceedings of the Advances in neural information processing systems, pp. 1097–1105, Long Beach, CA, USA, December 2012.

[10] F. Chollet, “Xception: deep learning with depthwise separable convolutions,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1251–1258, Honolulu, HI, USA, July 2017.

[11] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size,” 2016, https://arxiv.org/abs/1602.07360.

[12]Priyanka Malhotra, Sheifali Gupta, Deepika Koundal, Atef Zaguia, Wegayehu Enbeyle, "Deep Neural Networks for Medical Image Segmentation", Journal of Healthcare Engineering, vol. 2022, Article ID 9580991, 15 pages, 2022. https://doi.org/10.1155/2022/9580991

[13] W. Liu, A. Rabinovich, and A. C. Berg, “Parsenet: looking wider to see better,” 2015, https://arxiv.org/abs/1506.04579.

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). *Attention is all you need*. Ithaca: Cornell University Library, arXiv.org.

[15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., . . . Houlsby, N. (2021). *An image is worth 16x16 words: Transformers for image recognition at scale*. Ithaca: Cornell University Library, arXiv.org.

[16] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems* 27 (2014): 3104-3112.

[17]Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv, abs/2010.11929.

[18]Mian Wu, Yinling Qian, Xiangyun Liao, Qiong Wang, and Pheng-Ann Heng. Hepatic vessel segmentation based on 3dswin-transformer with inductive biased multi-head self-attention. arXiv preprint arXiv:2111.03368, 2021.

[19]Shamshad, F. , Khan, S. , Zamir, S. W. , Khan, M. H. , Hayat, M. , & Khan, F. S. , et al. (2022). Transformers in medical imaging: a survey. arXiv e-prints.

[20] Raghu M , Unterthiner T, Kornblith S , et al. Do Vision Transformers See Like Convolutional Neural Networks?[M]. 2021.

[21] Lin, T., Wang, Y., Liu, X., & Qiu, X. (2021). *A survey of transformers*. Ithaca: Cornell University Library, arXiv.org.

[22] Sun, Z., Cao, S., Yang, Y., & Kitani, K. (2021). *Rethinking transformer-based set prediction for object detection*. Ithaca: Cornell University Library, arXiv.org.

Wang Ru-yi, bone in 2003, none

.

Wang Ren-li， born in 2003, none

[键入文字]