南京工業大學

2022 届毕业设计（论文）

题　　目：基于深度学习的早产儿眼底图像黄斑区及视盘检测

专　　业：计算机科学与技术（嵌入式）

班　　级：计（嵌入）1802 班

姓　　名：林瀚

指导老师：吴梦麟

起讫日期：2021.12-2022.6

2022 年　6　月

基于深度学习的早产儿眼底图像黄斑区及视盘检测

摘　要

黄斑区及视盘的目标检测在早产儿视网膜病变的诊断中发挥着重要作用，可用于精确定位和病变分区等任务上。过去的传统图像处理算法存在许多问题。而近几年深度学习的火热和神经网络在目标检测上的应用，克服了传统算法的限制，让机器自主学习的方法代替人工提取眼底图像特征，并通过优化网络结构显著提高了检测精度和速度，大力推动了黄斑区及视盘定位任务的发展。

本文提出了一种改进 Mask R-CNN：在损失函数中加入距离限制的同时，改进全连接头部为双头结构。一方面，在深度神经网络模型训练时加入距离监督可以弥补黄斑区和视盘同时定位造成的精度损失，而且能够作为动态学习率加快模型收敛；另一方面，通过结合全连接头的空间敏感性和卷积头的空间相关性，改进 R-CNN 网络的检测头部结构能够进一步提高模型分类和定位的精度。实验通过消融实验证明了距离监督的有效性，与无距离监督方法相比，本文的方法得到了更高的精度。

关键词 ：Mask R-CNN 深度学习 目标检测 黄斑定位 视盘定位

Macula and Optic Disc Detection of Retinal Fundus Images in Premature Infants

Abstract

Object detection of macula and optic disc plays an important role in the diagnosis of retinopathy of prematurity, which can be used in tasks such as precise positioning and lesions partition. There are many problems in the past image processing algorithms. But deep learning is on the rise and the neural networks implemented on object detection, overcoming the limitations of traditional algorithms, allow machine to learn the fundus image features autonomously instead of extracting them artificially. It has significantly improved detection accuracy and speed by optimizing network structure, vigorously promoting the development of macular and optic disc positioning.

This paper proposes an improved Mask R-CNN: while adding distance supervision to the loss function, the fc-head is improved to a double-head structure. On the one hand, distance supervision can make up for the loss of accuracy caused by the simultaneous positioning of the macula and optic disc, and can also be used as a dynamic learning rate to speed up the convergence of model; on the other hand, by combining the spatial sensitivity and the spatial correlation of the two heads, the improved detection head structure can further improve the accuracy of classification and localization. Experiments demonstrate the effectiveness of distance supervision through ablation experiments, and our method achieves higher accuracy compared to methods without distance supervision.

Keywords: Mask R-CNN; Deep learning; Object Detection; Macular Detection; Optic Disc Detection

目　录

第一章 引言

1.1 背景

1.1.1 眼底结构

正常的眼底结构由黄斑区（Macula）、视盘（Optic Disc）和大量视网膜血管组成。

黄斑区在视网膜眼底图像中表现为暗红色圆斑，是视力最敏锐的部位。因此，在此处发生的任何病变都会对视力造成严重的伤害。通常位于视网膜血管最为密集的中心区域，同时不包含任何血管。

视盘在视网膜眼底图像中表现为境界分明的淡红色圆盘，是视网膜上视觉纤维汇集的部位，通常位于视网膜由黄斑向鼻侧方向约 3mm 处。

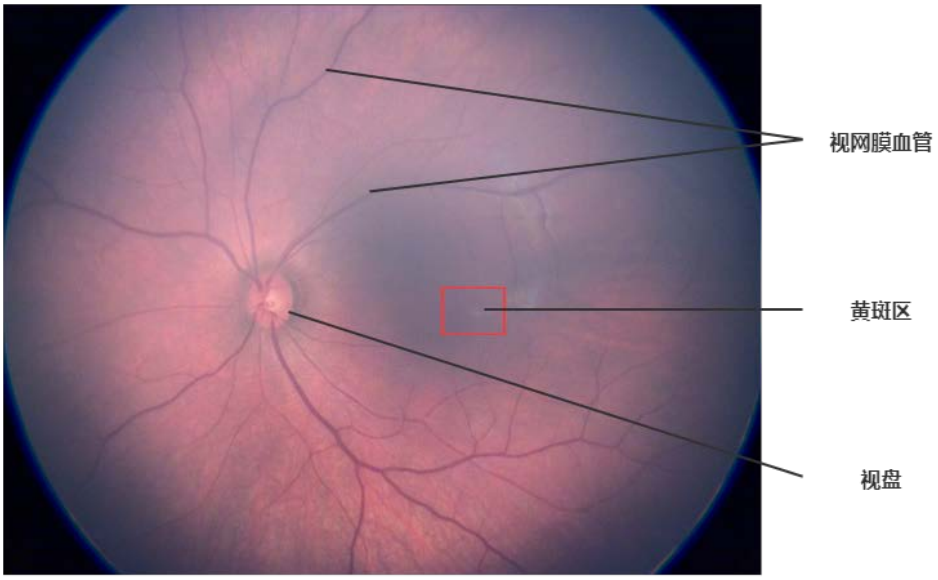


图 1-1 视网膜图像中的眼底结构

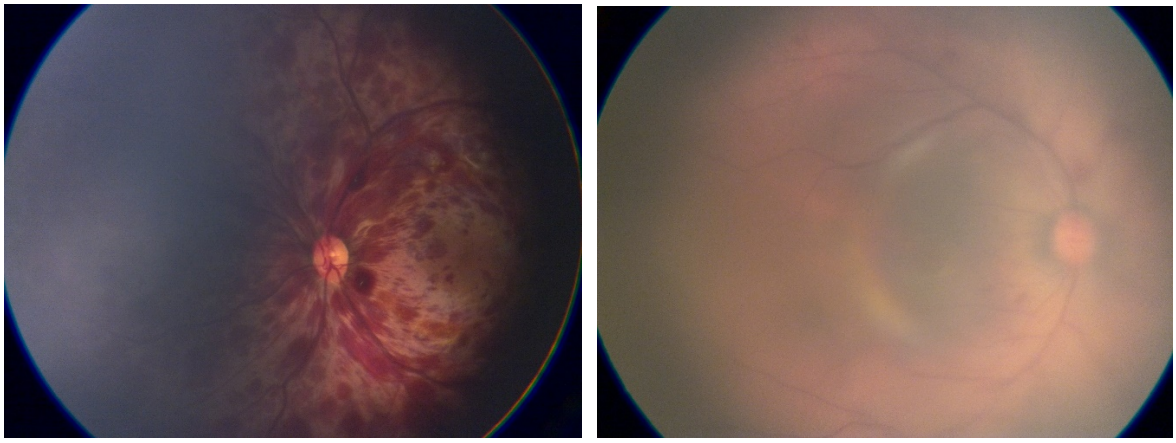
1.1.2 早产儿视网膜病变

早产儿视网膜病变（ROP），是由于早产儿在视网膜血管发育时期接触了高浓度氧气，未完全血管化的视网膜发生血管收缩和增殖以及由此产生导致的。关于病变的分区标准[6]，一般会将视网膜分为三个区：其中以视盘为中心，视盘到黄斑区中心的两倍距离为半径的圆形区域称为Ⅰ区；向外直到鼻侧视网膜周边部划圆，该环形区域称为Ⅱ区；Ⅱ区以外剩

留的颞侧半月形区称为III区。早产儿视网膜病变严重程度，从 I 区到III区逐级递减。主治医生主要是通过对早产儿视网膜发育程度和病变所在分区进行分析，从而判断早产儿的病情并决定后续的治疗手段。

1.2 研究目的及意义

在发生视网膜出血、视网膜周边白斑等病变，或者拍摄角度和光线等外部因素影响下，早产儿视网膜眼底图像会如图 1-2 (a)，(b)所示，难以确认黄斑区位置。若是人工进行定位，因为视盘边界清晰反光较强，最好的办法是首先确认视盘的位置，然后大致通过视盘和黄斑区的距离关系推测出黄斑区的位置。但是这样不仅精度低，效率差，而且具有很大的主观性和经验性，部分专业能力较弱的医生会出现漏检或者误判的情况。再者，当拍摄的视网膜图像中不存在视盘时，就只能通过其他方法，如寻找视网膜血管汇集处等对黄斑区进行定位。不过，这些方法在图像质量较低的情况下，依然存在着较大的偏差。



（a） （b）

图 1-2 难以辨认的视网膜图像：(a)病变，(b)光线不足

因此该研究旨在使用计算机技术对质量较低，尤其是发生大量病变的视网膜图像，进行视盘和黄斑区精确且快速的目标检测和定位。准确的视盘和黄斑区位置坐标，不仅可以减轻医生的判断和定位压力，无需通过肉眼辨别，而且能够用来计算出视盘和黄斑区之间的距离大小，结合中心点坐标为后续的病变分区提供辅助，大幅增加了诊断效率。

1.3 研究现状

虽然目前关于早产儿视网膜目标检测的研究较少，但是对于成人视网膜眼底图像的黄斑区及视盘目标检测技术已经较为成熟，而且早产儿和成人的视网膜除了发育程度不同其余结构完全一致。通过阅读文献发现，技术方向主要分为传统计算机图像处理算法和近年来流行的深度学习算法。

1.3.1 传统图像处理算法

传统算法主要包括[1]基于血管提取分割的定位、应用方向局部对比度滤波结合局部血管密度[10]方法、阈值法、模板匹配法等。这些方法主要思路为提取几何形状、血管等特征，或是依赖视盘、黄斑区和血管之间的信息关系，根据这些特征信息实现黄斑区和视盘的定位。

1.3.2 深度学习算法

传统定位算法存在以下问题：

需要将视盘和黄斑区分步骤处理，时间耗费长而且精度较差。传统算法的人工特征提取具有很强的主观性和经验性，这会对模型训练带来一些影响。同时在调节模型时,也需要耗费大量的时间和精力。相比视盘，黄斑区的边界不清晰且反光较弱，因此传统算法对黄斑区的定位效果较差。加上病变、拍摄光照和对比度等因素对视网膜眼底图像的影响，依赖人工特征提取的传统算法有着很大的局限性。

而深度学习算法克服了上述问题，通过让机器自主学习的方法，提取数据原始特征；端到端的学习只需调节超参数就可对模型进行训练拟合。不仅可以通过多任务，实现视盘和黄斑区的同时检测，而且利用 GPU 加速神经网络，可以使用更短的时间达到更高的精度。

目前常用的目标检测深度学习算法主要分为两类：

单阶段算法（one-stage），如 YOLO[2]、SSD[3]、Retina-Net、DetectNet、SqueezeDet 等，通过卷积神经网络提取特征，直接预测目标的分类与定位。相比双阶段算法而言，速度较快而且能够避免背景错误而导致的误判，但是精度较低，对于较小物体的检测效果较差。

双阶段（two-stage）算法，如 Faster-RCNN[7],[12]，Mask-RCNN[11]，SPPNet，R-FCN 等，通过训练一个 RPN 候选区域生成网络，为卷积神经网络提取出来的图像特征生成对应的候选区域框，再通过卷积神经网络预测目标的分类和定位。该算法利用了 Anchor 机制，精度较高，同时通过 FPN 架构可以有效解决不同大小的物体检测问题，缺点是速度较慢，训练时间长。

第二章 相关工作和背景介绍

考虑到视盘和黄斑区的尺寸较小且追求较高的精度，本文提出了一种基于 Mask R-CNN[11]的改进算法。本章中首先将介绍 Mask R-CNN[11]的原理和结构,它是目前较常用的双阶段算法之一。接着将阐述双头检测头结构的背景原理等，作为下一章关于检测头改进的铺垫。

2.1 Mask R-CNN

2.1.1 简介

Mask R-CNN[11]主要灵感来自于 Faster R-CNN[12]，沿用了以 RPN 为第一阶段的双阶段结构，修改了头部结构可以同时生成物体分类、定位和掩膜（mask），可用于实例分割和目标检测领域。不仅如此，它还结合 FPN 结构改进了初始的特征提取网络，实现了物体多尺度检测；使用 RoI Align 代替了原 RoI Pooling 来提高精度。

基本架构大致如图 2-1 所示：

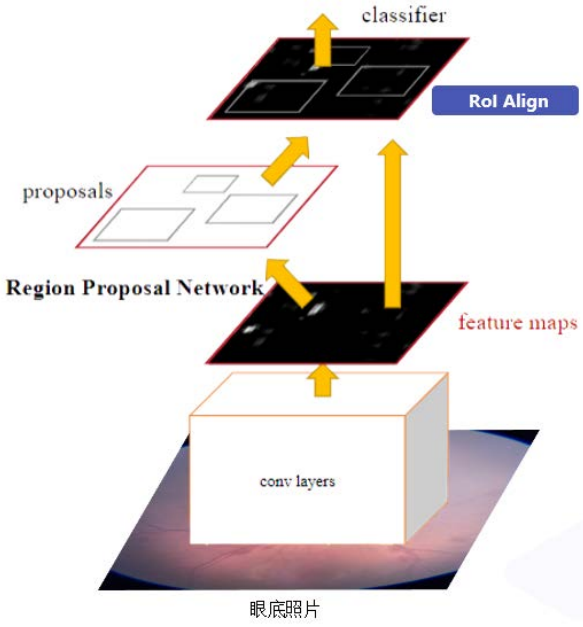


图 2-1 Mask R-CNN 的基本架构

2.1.2 原理

特征提取网络（conv layers）：

本文使用了主要由卷积层、池化层、ReLu 激活层、Batch Norm 层和残差连接组成的残差神经网络作为特征提取网络。作为滤波器的卷积层可以通过移动卷积核进行卷积运算提取图片的视觉特征如边界、颜色等，另一方面，权重通过卷积核共享，降低了参数的数量；池化层的作用是降低数据的空间大小和参数量，从而减少计算资源的耗费，也能有效抑制过拟合；激活层的 ReLu 函数使得梯度下降更为稳定，模型非线性；Batch Norm 层的目的是减轻数据分布经过神经网络传递后的变化，用可训练的参数对输出数据进行归一化，以此增加模型的训练速度和泛化能力；作为残差神经网络核心的残差连接，和卷积层、Batch Norm 层及激活层构成了如图 2-2 所示的残差块。深度神经网络在进行前向传播时，随着网络深度不断加深，传递的信息会不断递减，这将会导致神经网络发生梯度消失、网络退化等问题。而残差连接通过传递映射与残差之和，使得网络前后传播时不仅能够始终保留原始信息，并且增加了训练中获取的新信息，一定程度上缓解了上述问题。

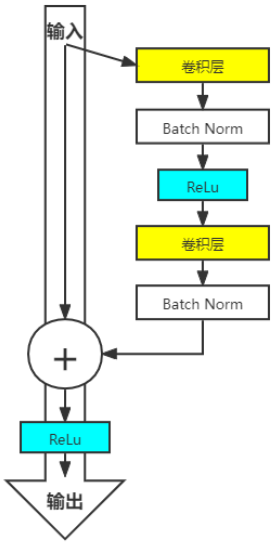


图 2-2 残差块的基本结构

FPN（Feature Pyramid Network）：

FPN 是一种通用多尺度架构，经常出现在主流的目标检测方法中，可以结合各类骨干网络使用。FPN 结构主要由自下而上、自上而下和横向三个连接构成：自下而上和自上而下的连接分别指的是经过卷积神经网络的下采样和上采样，横向连接指的是相同尺度大小

特征图的融合。这种通过连接实现的特征融合，赋予了输出的特征图很强的语义信息和空间信息。

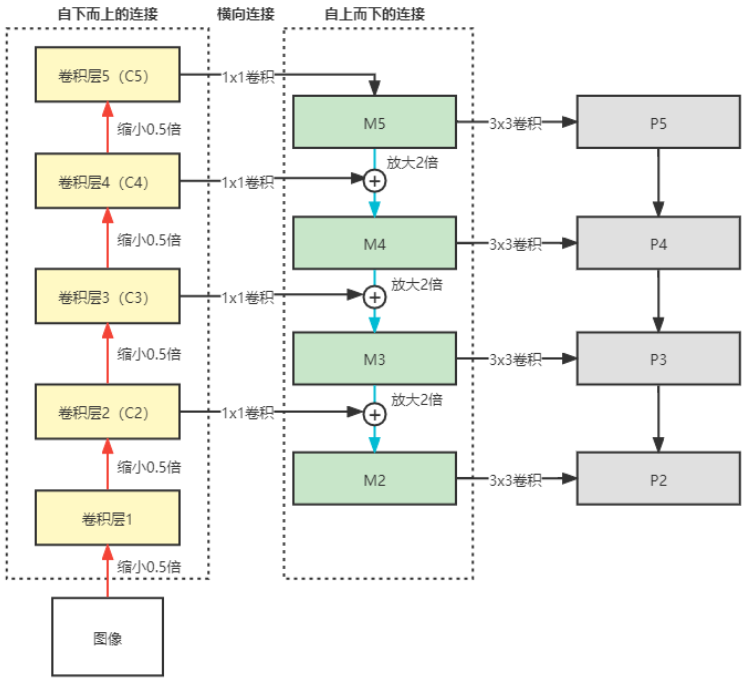


图 2-3 FPN 结构

提取出尺寸大小分别为原图大小 $\frac{1}{4}$ ，$\frac{1}{8}$ ， $\frac{1}{16}$ ，$\frac{1}{32}$ 的四张特征图

传统 Faster R-CNN[12]只使用了特征提取网络的最后一层特征图，然而实际上对于小目标来说，当卷积池化进行到最后一层时，语义信息已经不复存在了，这会导致检测小物体的能力较差。而 Mask R-CNN[11]不仅在前述特征提取网络中加入了 FPN 结构来提取多张不同分辨率的特征图，并在后续 RPN 中使用多层特征图与 anchor 机制结合，生成多尺度的候选区域。

RPN（Region Proposals Networks）：

与经典目标检测算法的滑动窗口不同，RPN 能够直接生成可能存在目标的候选框，极大提升检测框的生成速度。当接收到特征提取网络传递的数张特征图时，RPN 内部的锚点（anchor）机制会为每张特征图的每个像素点生成多种锚框，例如本文使用的 12 种锚框：面积分别为 32×32、64×64、128×128、256×256，长宽比分别为 1:1、1:2、2:1，面积和长宽比两两组合。不同分辨率的特征图与多样的锚框相配合，解决了网络的多尺度检测问题。

之后再通过数层卷积层，返回每个像素点的目标分类分数和边界坐标回归，结合锚框生成候选区域。其中目标分类为前景和背景的二分类。

因为进行目标检测的过程中，在同一目标的位置上会产生大量的候选区域，而这些候选区域之间可能会存在重叠。为了解决该问题，RPN 会对输出结果中分类分数较高的候选区域进行非极大值抑制（NMS）以去除单个目标的重复目标框。非极大值抑制的算法流程大致如下：1. 根据置信度进行排序 2. 选择置信度最高的边界框，将其添加到最终输出列表中 3. 计算置信度最高的边界框与其他候选区域的交并比（IoU）4. 删除交并比大于设置阈值的边界框 5. 重复上述过程，直到边界框的列表为空。

最后，再汇总多张特征图的结果，挑选出置信度较高的数个候选区域传递给网络的下一部分。

RoIAlign（Region of Interest Align）：

RoIAlign[11]是对 RoI Pooling[12]的改进。因为传统的卷积层的输入与输出都是固定尺寸，所以它们的目的都是对大小不一的候选区域中的特征进行相同大小的统一表示。为了实现这一想法，RoI Pooling[12]加入了两次量化：当候选区域边界位于特征图的单元像素之间时，进行的取整操作；当不同大小的候选区域需要经过池化获得相同大小的特征图时，对无法直接均分的子区域进行的取整操作。但经过两次浮点数直接取整，带来的像素偏差显然会对后续的分类和定位产生一些影响。于是，RoIAlign[11]将取整操作改进为双线性插值法，减少了量化所带来的误差，在小物体目标检测上更为精确。

检测头（Detection Head）：

经过 RoIAlign 的候选区域（边界坐标以及对应的特征图）最后会被传递到网络的检测头部中，进行目标的分类和第二次边界回归。此处产生的损失会同 RPN 中的损失一起进行反向传播，从而进行端到端的训练。

2.2 基于双头结构的检测头

原 Mask R-CNN[11]仅使用了由双层全连接层构成的头部结构来进行目标分类和边界回归。通过查阅文献发现双头结构[13]在这两个任务上表现更加优秀。

2.2.1 背景与原理

绝大多数的双阶段目标检测算法都是共享单个头部进行目标分类和边界框回归，而且头部结构一般使用全连接头（fc-head）或者是卷积头（conv-head）。例如 Mask R-CNN[11] 中的检测头结构是由双层全连接层组成，并同时进行目标分类任务和定位任务。然而实际上，这两种头部结构是互补的[13]：全连接头具有空间敏感性，比起卷积头，在候选区域和真实框上的交并比与分类分数的相关性更强，更适用于分类任务；而卷积头具有空间相关性，回归框回归更加准确，在定位任务上表现更好。

这一现象来源于两者的结构差异：全连接头在输入特征图的不同位置应用非共享参数的变换，隐含了空间信息，虽有助于区别完整物体和部分物体，但是对于整个物体的目标框的偏移量回归并不鲁棒；相比之下，卷积头在输入特征图的所有位置上使用共享参数的变换，即卷积核，并使用池化层对空间信息进行聚合，空间之间的相关性更强。因此，结合两者的优势和特点，利用全连接头和卷积头分别进行分类和定位任务能够有效提升模型的精度。

2.2.2 普通双头结构

基于上述结论，原文提出了一种双头方法（Double-Head）来利用这两种检测头结构的优势：它有一个用于分类的全连接头（fc-head）和一个用于边界框回归的卷积头（conv-head），并让每个头部结构都专注于其被分配的任务。

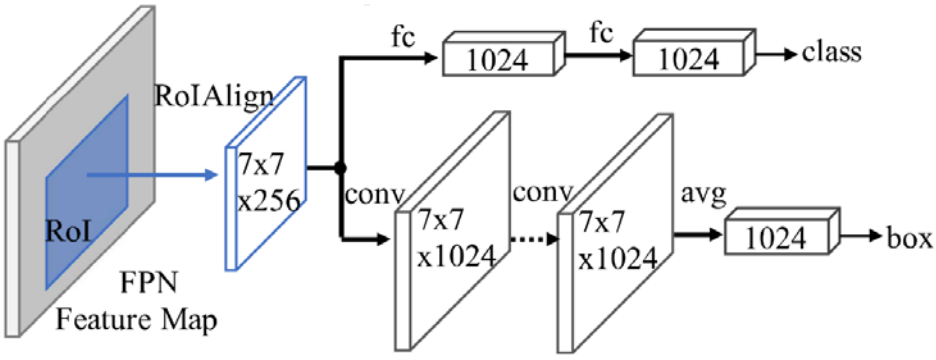


图 2-4 Double-Head[13]的基本结构

经过 RoIAlign 的特征图被同时传给全连接头和卷积头。全连接头使用两层全连接层；卷积头使用多个卷积模块的堆叠和平均池化层。

在此基础上，原文发现通过对普通双头结构进行拓展，引入非专注任务（Leveraging Unfocused Tasks）[13]能够进一步提升双头结构的效果。因此本文将借鉴该双头结构的拓展作为检测头的主要改进方向，将于下一章中详细介绍。

第三章 基于双头结构和距离约束的黄斑区及视盘检测

本章中将详细介绍本文提出的改进 Mask R-CNN，主要包括针对原 Mask R-CNN[11]全连接检测头的双头拓展改进，以及根据黄斑区与视盘距离信息提出的距离约束。最后将介绍该模型的具体实现和相关细节。

3.1 双头结构拓展

3.1.1 简介

本文将原 Mask R-CNN[11]的全连接头部改良为如图 3-1 所示的 Double-Head-Ext[13]结构。与单纯地把分类任务和定位任务分别交给全连接头和卷积头的普通双头结构不同，引入了非专注任务。这么做是因为通过实验发现全连接头中的边界框回归任务能够反过来对分类提供辅助监督，而且由于两个头截然不同的结构能够为目标分类捕获互补的信息。配合原文提出的一种互补的融合方法，将两个检测头的分类器进行融合之后分类效果更好。

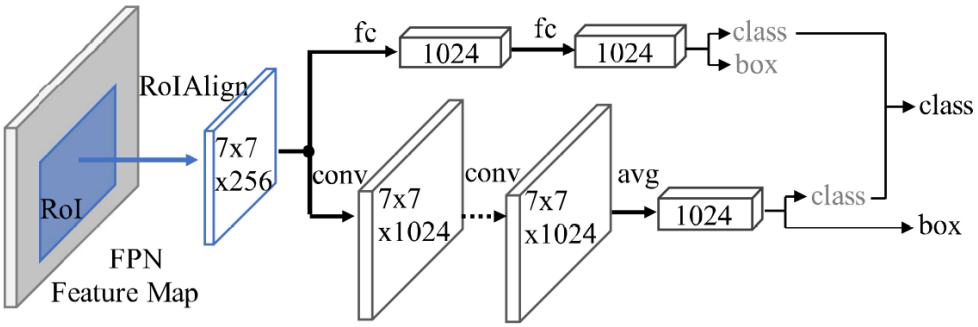


图 3-1 Double-Head-Ext[13]的基本结构　前端与普通双头结构相同，主要在检测头后端进行拓展，两头都输出目标分类和边界框进行非专注任务。

3.1.2 细节描述

分支结构：

全连接头分支的结构依然还是使用原来的双层全连接层；关于新加入的卷积头分支，本文遵循原 Double-Head-Ext[13]中的设计，使用了由图 3-2 所示的残差模块、瓶颈模块（Bottleneck Block）和非局部模块（Non-local Block）的堆叠，并采用了实验效果较好的 1 个残差模块+2 个瓶颈模块+2 个非局部模块的组合。第一个残差模块主要作用是增加通道

数到 1024，之后在每个瓶颈模块之前插入一个非局部模块来获取全局信息，以增强前景对象的信息。每个卷积层之后都有一个 Batch Norm 层。

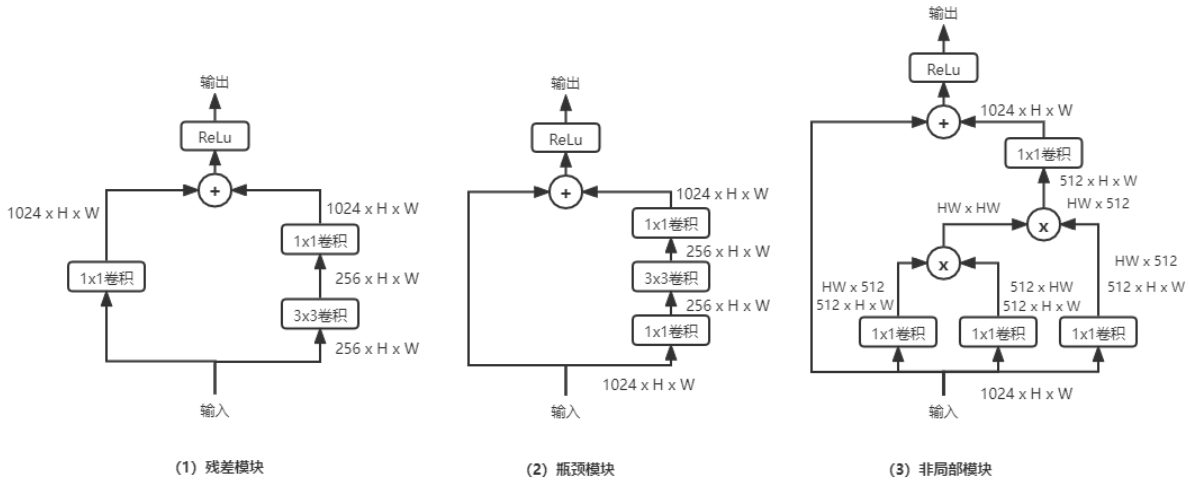


图 3-2 三种模块结构：（1）增加通道数的残差模块（2）瓶颈模块

（3）非局部模块；每层卷积层后隐含一个 Batch Norm 层。

分类融合：

分类融合方法使用了原文中较为推荐的互补融合（最大和平均），如下所示：

$$s = s^{fc} + s^{conv}(1 - s^{fc}) = s^{conv} + s^{fc}(1 - s^{conv}), \quad (1)$$

其中 s^{fc}，s^{conv} 分别为全连接头分支和卷积头分支的分类器结果。

损失函数：

由于引入了非专注任务，损失函数也相应有所优化。将在 3.3.1 节中与距离约束一同详细阐述。

3.2 距离约束

显然通过大量的训练，卷积神经网络能够学习到目标的结构特征。但是除了黄斑区和视盘的结构特征，还有“视盘大致位于视网膜由黄斑向鼻侧方向约 3mm 处”这一隐含距离信息没有利用。然而早产儿的视网膜存在发育不完全和病变的情况，为了确认这一距离信息是否也适用于于早产儿，本文通过对 2464 张同时存在黄斑区和视盘的早产儿视网膜

眼底图像进行距离测量，发现整体距离分布呈以 800 像素距离为均值的正态分布（图 3-3）。这从一定程度说明了无论是对于成人还是早产儿，黄斑区与视盘之间确实存在一定范围的距离约束。

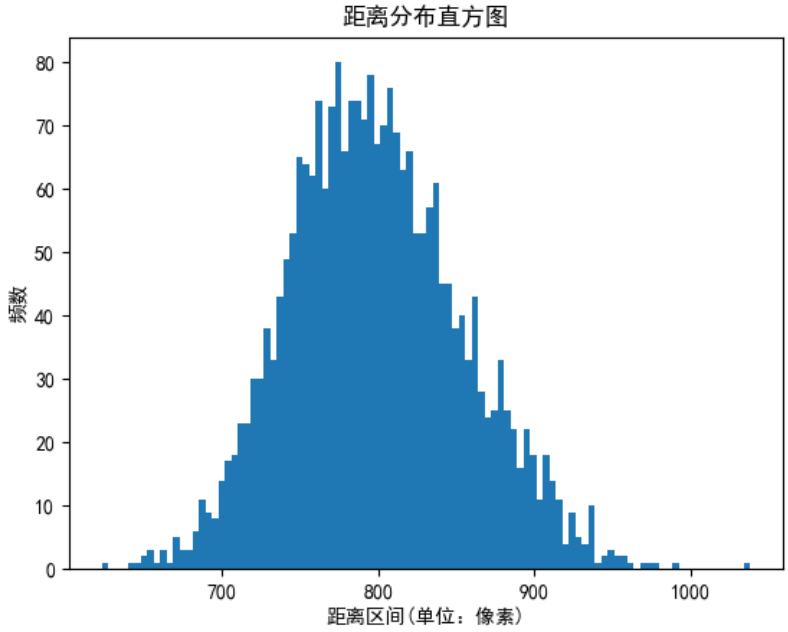


图 3-3 黄斑区和视盘之间的欧式距离分布图。整体曲线呈现以 800 为均值的正态分布

由此，本文提出一个设想：如果实际上网络只能学习到黄斑区和视盘的结构特征，而并不能从数据中学习到隐含的距离信息的话，可以考虑在模型训练中加入距离监督，强制网络对生成的黄斑区及视盘定位之间的距离进行约束，从而达到提高精度的效果。

根据上述分析，本文在训练中引入了距离监督，在黄斑区和视盘之间加上一定距离约束来让网络能够学习到距离的深层信息。不过，因为每位早产儿的发育和先天条件不同，再加上拍摄时的角度影响，将距离完全限制在一个固定值或是一个较为精确的区间上反而会影响精度。另外，希望当黄斑区和视盘距离过近或过远时能够加速收敛。于是取中间 95% 置信区间作为标准：当距离落在区间内时，损失为 0；当距离落在区间外时，损失以指数函数增长。距离损失 $\mathcal{L}^D$ 定义如下：

$$\mathcal{L}^D(x)=\begin{cases} e^{\frac{MIN-x}{SCALE}}-1, & x<MIN \\ 0, & MIN\le x\le MAX \\ e^{\frac{x-MAX}{SCALE}}-1, & x>MAX \end{cases} \tag{2}$$

其中，MIN、MAX分别为置信区间的左右端点，$SCALE$为指数的缩放倍数，x为黄斑区和视盘回归边界框的中心点距离差。将选取黄斑区和视盘分类分数最高且超过一定阈值的边界框来进行距离计算，低于相应阈值则认为不存在。该损失仅在黄斑区和视盘同时存在的情况下生效。

3.3 具体实现

3.3.1 非专注任务与距离约束下的损失函数

原来的损失函数由 RPN 和全连接检测头的损失组成，包括各自的分类交叉熵损失和边界回归 Smooth L1 损失。改良后的损失函数如下所示：

$$\mathcal{L} = w^{fc}\mathcal{L}^{fc} + w^{conv}\mathcal{L}^{conv} + \mathcal{L}_{cls}^{RPN} + \mathcal{L}_{reg}^{RPN} + w^{D}\mathcal{L}^{D} \tag{3}$$

其中，w^{fc}，w^{conv}，w^{D}分别是全连接头分支，卷积头分支和距离限制的权重，$\mathcal{L}^{fc}$，$\mathcal{L}^{conv}$，$\mathcal{L}_{cls}^{RPN}$，$\mathcal{L}_{reg}^{RPN}$，$\mathcal{L}^{D}$依次分别为全连接头分支损失，卷积头分支损失，RPN 分类损失，RPN 边界回归损失和距离损失。

非专注任务：

由于引入了非专注任务，需要分别在全连接头分支和卷积头分支中加入一个权重，来平衡分类损失和边界回归损失。全连接头分支损失原文定义如下[13]：

$$\mathcal{L}^{fc} = \lambda^{fc}L_{cls}^{fc} + (1-\lambda^{fc})L_{reg}^{fc}, \tag{4}$$

其中，L_{cls}^{fc} 和 L_{reg}^{fc} 分别是全连接头中的分类和边界回归损失，λ^{fc} 是控制全连接头中两个损失之间平衡的权重。类似的，卷积头分支损失定义如下[13]：

$$\mathcal{L}^{conv} = (1-\lambda^{conv})L_{cls}^{conv} + \lambda^{conv}L_{reg}^{conv}, \tag{5}$$

其中，L_{cls}^{conv} 和 L_{reg}^{conv}分别是卷积头中的分类和边界回归损失。不同于公式（3）中 λ^{fc} 乘以分类损失 L_{cls}^{fc}，卷积头中的平衡权重 λ^{conv} 乘的是边界损失 L_{reg}^{conv}，这是因为定位才是卷积头的重点任务，而不是目标分类。

距离约束：

$\mathcal{L}^D$ 使用公式（2）的定义进行计算，并在距离损失中加入一个权重 w^D。一是为了与双头结构、RPN 网络产生的损失相平衡。二是因为在实际训练中，由于指数函数的特性常引起梯度爆炸从而模型无法收敛的问题，需要一个较小的权重来对距离损失进行平衡。

3.3.2 网络架构与实现细节

接下来将结合前述改进，介绍本文提出的改进 Mask R-CNN 架构和实现细节。整体网络架构如图 3-4 所示。网络部分主要由特征提取网络、RPN 和 RoI Head 构成，输出部分包含 RPN 损失、距离损失、头部损失以及最终的检测结果。RPN_loss 包含前后景分类损失和边界回归损失；detector_loss 包含黄斑视盘分类损失和边界回归损失；distance_loss 为黄斑区和视盘的距离损失。

输入的图片数据首先需要经过缩放、标准化等预处理。原始图像的 3264 x 2448 分辨率均会被缩放到三分之一大小，即 1088 x 816，标准化使用的均值与误差通过数据集实际计算得到。预处理结束后进入特征提取网络，由 FPN 结构生成四张不同分辨率的特征图 P2，P3，P4，P5（图 2-3）；RPN 将利用这些特征图生成候选区域并传递给由 RoI Align 和检测头组成的 RoI Head；最后由 RoI Head 生成边界框和其对应的分类分数。

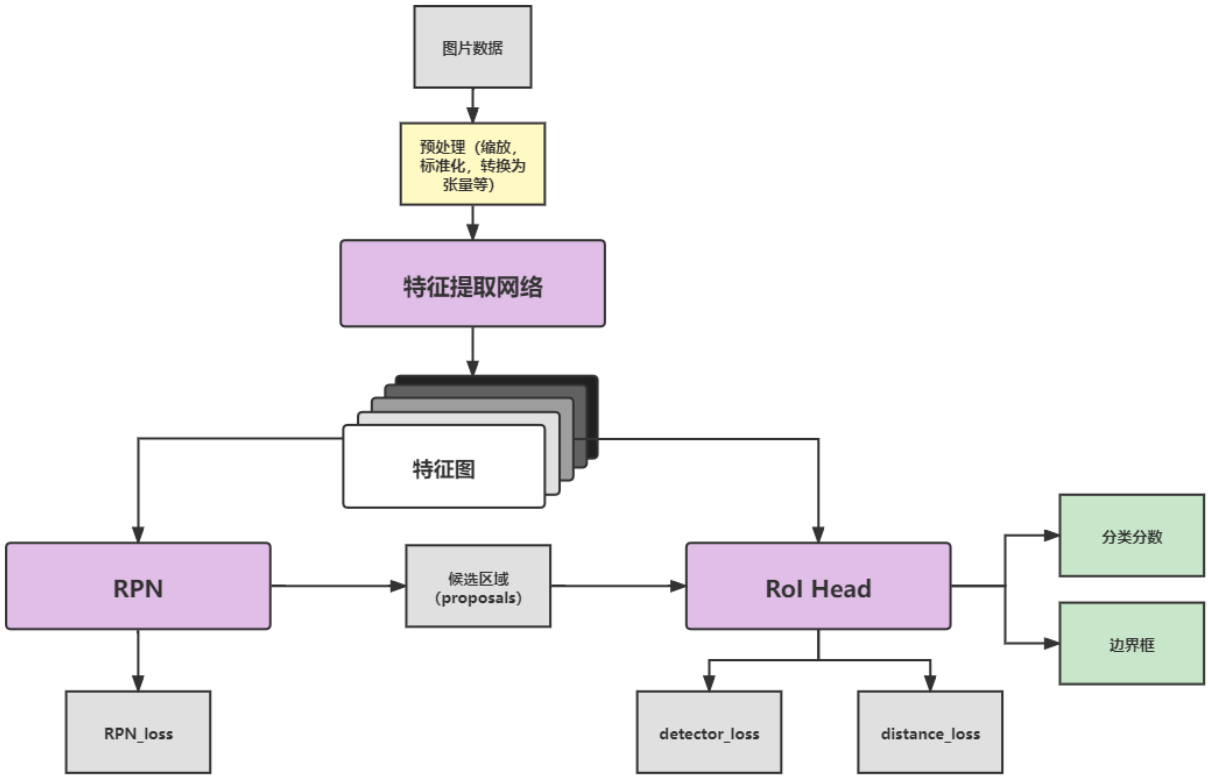


图 3-4 总体架构

其中，RPN 架构如图 3-5 所示。特征图经过 3x3 卷积层后被分为两个分支，并再次使用 1x1 卷积操作返回对应的边界框和分类分数。Anchor 生成器针对不同分辨率的特征图生成不同大小的多长宽比锚框（1:1，1:2，2:1），结合边界框和分类分数，经过非极大值抑制后按置信度选出最优的候选区域。

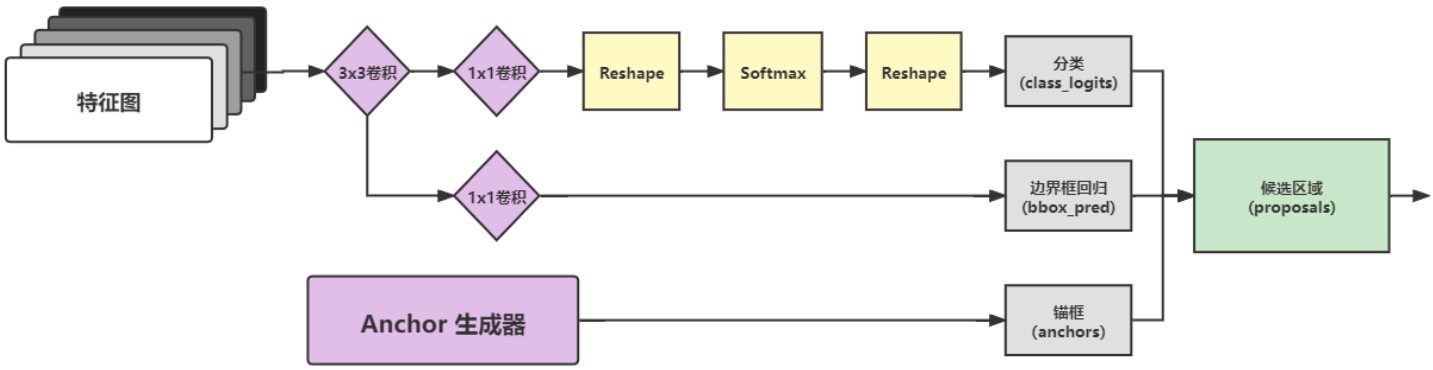


图 3-5 RPN 架构　分类分支主要判断物体是否属于前后景；Anchor 生成器针对 P2，P3，P4，P5，对应生成 32×32、64×64、128×128、256×256 面积大小的锚框

RoI Head 前端接收上层 RPN 传递的候选区域并再次复用特征图后，将它们传递给 RoI Align，生成大小通道为 7x7x256 的 RoI 特征图（图 3-6）。最后由 2.2 节介绍的 Double-Head-Ext[13]双头结构进行最终的结果输出：全连接头分支需要先将特征图转换为一维张量，再传入两层特征表示大小为 1024 的全连接层；卷积头分支中，特征图经过残差模块提升通道后传入两次非局部模块和瓶颈模块的堆叠，返回 7x7x1024 的三维张量，再由池化层将结果变形为和全连接头分支输出相同的 1024 维向量。两条分支输出的 1024 维向量将会再次经过特征表示分别为 3（分类类别：背景、黄斑区、视盘）和 3*4（分类类别*边界框四点坐标）的全连接层，得到最后的结果。

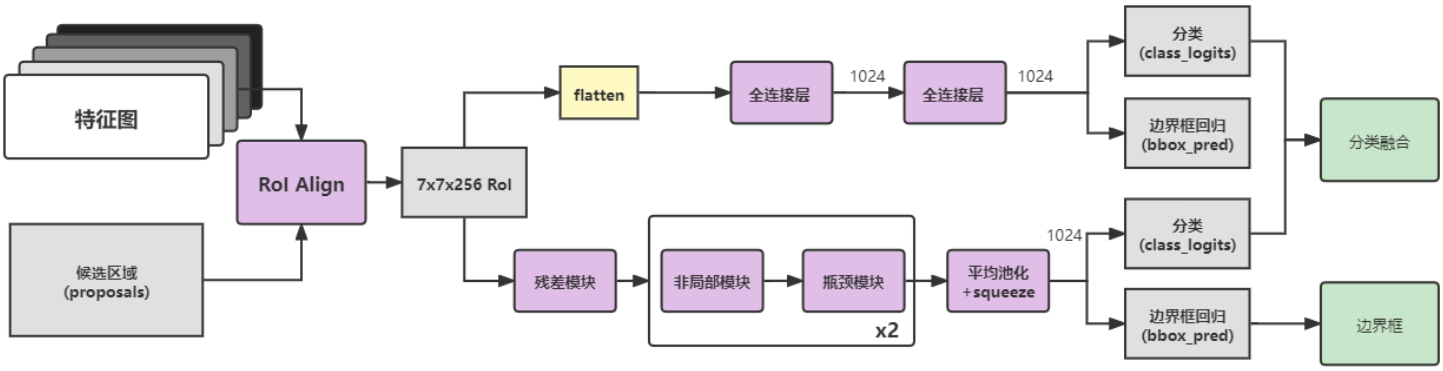


图 3-6 RoI Head 架构　flatten 和 squeeze 作用为将多维张量转换为向量。

至此，本文提出的改进 Mask R-CNN 模型已介绍完毕。在下一个章节中，将通过实验验证该模型的有效性。

第四章 实验结果

本章中将在自制数据集上评估本文的改进 Mask R-CNN 模型，并通过消融实验来对模型进行分析。

4.1 数据来源与标签制作

目前可搜集到的公共集中，多为成人的视网膜眼底图像。为了实验的严谨性，使用了实地拍摄的数百名早产儿眼底图像，且由专业医生为我们标注了大量黄斑区的位置。因为视盘较为清晰，定位比较简单，可以自己通过 Labelme 软件进行视盘的标注。鉴于实际数据中存在大量病变或拍摄光线不足等情况，不再进行数据增广或噪声的添加。经数据清洗和随机抽样后，得到了一个共 5000 张眼底图像的数据集（4000 张作为训练集，1000 张作为测试集）。

然而在后续的实验中，观察到黄斑区的分类精度和定位效果与视盘差距巨大。通过检查错误样本发现，大量错误来源于模型的误判（False Positive）。结合实际标签对比发现，半数以上的误判样本是由于图像过糊或医生疏忽引起的黄斑区未标注，因此导致了对于模型的评估不够准确，而且该噪声也对模型本身产生了一定影响。去除所有黄斑区的负样本后，虽然问题得以解决，但是这么做会大大减少数据量，实验结果会变得较不稳定。于是在接下来的实验中，将采用所有样本进行消融实验。待选出最优参数后，再去除黄斑区的负样本进行模型的训练和评估。

4.2 实验流程与细节

实验流程简单分为数据读取、模型训练和模型评估三个环节。

数据读取：

在使用 Labelme 制作数据集时，将所有眼底图像的标签 json 文件都汇总在同一个 csv 文件中，以便后续使用 Pandas 读取。同时继承了 Pytorch 框架下的 Dataset 类，配合 DataLoader 进行数据的批量读取和加速。

模型训练：

所有模型都使用一块 NVIDIA RTX3090 和 32GB 内存进行训练。数据集中的图像均为 3264×2448 像素的实拍眼底照片。使用的超参数和技巧如下：训练轮次（epochs）为 20 个轮次；初始学习率设为 1e-4，使用带有 Warmup 的余弦退火学习率衰减；优化器选择 Adam，参数使用 Pytorch 下的默认参数；损失函数中根据经验将w^{fc}，w^{conv}，w^{D}分别设置为 2.0，2.5，0.5；λ^{fc}，λ^{conv}，$SCALE$分别设置为 0.7，0.8，1000；图片标准化使用的均值和方差由实际计算得到，分别为 [0.462, 0.372, 0.386] 和 [0.182, 0.131, 0.119]；固定了随机数种子来减少随机性对实验的影响。特征提取网络的 FPN 骨干（backbone）使用预训练的 resnet18、resnet34、resnet50、resnext50_32x4d、resnet101、resnext101_32x8d 并冻结前两层残差层，RPN 和网络头部采用端到端联合训练。

模型评估：

每轮训练结束后，都会使用测试集图片进行模型的评估。评估标准主要使用平均交并比和各类别 AP（平均精确度）。其中，AP 的计算算法采用 VOC2012 年之后的算法。由于黄斑区结构的特殊性，在最终实验结果中还加入了中心距离差来衡量定位的精确程度。

4.3 消融实验

本文选用了骨干为 resnext50_32x4d 和 resnext101_32x8d 预训练模型的特征提取网络，在 20 个轮次和批量大小为 16、8 的条件下进行一些消融实验来分析模型。

4.3.1 单任务与多任务的差距

- 单任务：黄斑区和视盘进行单独检测
- 多任务：黄斑区和视盘同时检测

此次实验使用精度较高的 resnext101_32x8d 骨干，检测性能如表 3-1 所示。为了减少数据集的特殊性（3.1 节）对实验的影响，黄斑区的检测效果仅使用平均交并比衡量，误判的样本并不会影响检测结果的交并比。

表 4-1 单任务检测与多任务检测的性能评估　特征提取网络是以 resnext101_32x8d 为骨干的 FPN。第一行显示了多任务的性能；第二行和第三行显示了黄斑区和视盘单任务的性能。

backbone = resnext101_32x8d	黄斑区平均交并比	视盘平均交并比	视盘 AP
黄斑区和视盘同时检测	0.72739±0.12009	0.87703±0.05953	0.984
单独检测黄斑区	0.72925±0.12556	\	\
单独检测	\	0.87961±0.05595	0.985

单任务的性能以略微的优势优于多任务，这是因为黄斑区和视盘之间特征存在差异，使网络在进行目标分类和定位时互相影响，难以兼顾彼此[1]。于是，本文认为模型仅仅关注到了黄斑区和视盘的结构特征，而并没有学习到它们之间的深层距离信息。为了验证这一结论，将会在下一小节进行距离限制的消融实验。

4.3.2 距离限制

此次实验使用训练速度较快的 resnext50_32x4d 骨干，并凭经验选取了多个关于距离损失函数的超参数，对于每对超参数（$SCALE$ 和 w^D）都训练一个模型来进行加入损失函数前后的模型性能比较。检测性能如表 3-2 所示。同 3.3.1 节，黄斑区的评估依旧只使用平均交并比。

表 4-2 多任务下加入距离限制前后的性能评估。特征提取网络是以 resnext50_32x4d 为骨干的 FPN。第一行基线显示了不加入距离限制的原性能；后数行显示了不同超参数下加入距离限制后的模型性能。当 $SCALE$ = 1000，w^D = 0.5 时，模型达到相对最优。

backbone = resnext50_32x4d	黄斑区平均交并比	视盘平均交并比	视盘 AP
基线	0.72399±0.12236	0.87573±0.05612	0.976
$SCALE$ = 300，w^D = 0.05	0.72326±0.12961	0.88154±0.05677	0.98
$SCALE$ =500，w^D = 0.5	0.72272±0.12489	0.87853±0.05564	0.983
$SCALE$ = 800，w^D = 0.5	0.71822±0.12767	0.87677±0.05404	0.983

$SCALE$ = 1000，w^D = 5	0.72583±0.11807	0.87946±0.05479	0.983
$SCALE$ = 1000，w^D = 0.5	**0.72533±0.12964**	**0.87299±0.05562**	**0.986**

加入损失函数后，黄斑区和视盘的检测效果都有些许提高，尤其是对视盘的 AP 有着不同大小程度的提升。当 $SCALE$ 取 1000，w^D取 0.5 时，提升最大。实验结果证明了我们加入的距离损失函数的有效性，同时验证了上一节中关于模型并没有学习到距离信息的猜想。实验中还发现，加入损失函数后模型收敛速度较之前快了近半个训练轮次，我们推测是设计的损失函数中的指数发挥作用，在模型刚开始训练的阶段精度较差且黄斑区和视盘之间距离差距过于极端，其作为动态学习率加速了模型收敛。但是，由于我们损失函数设计问题，当模型精度达到一定高度时，我们的结果会位于置信区间中间，导致该损失失效，从而提升的效果有限。未来可以通过改进损失函数的思路，对结果进行进一步的提升。

单任务的检测虽然精度高，但是检测速度较慢。而多任务的检测通过加入损失函数，可以弥补一部分的性能差距，同时提高检测速度。在本文的剩余部分中，我们将使用加入损失函数且超参数对为$SCALE$ = 1000，w^D = 0.5 的多任务检测模型表示改进 Mask R-CNN。

4.4 主要结果

本文在去除所有黄斑区的负样本的数据集上进行实验，并使用 resnet18、resnet34、resnet50、resnext50_32x4d、resnet101、resnext101_32x8d 这几个骨干对改进 Mask R-CNN 进行性能评估。表 4-3，4-4 显示了模型在黄斑和视盘检测上的精度。

表 4-3 不同骨干下的黄斑区精度

Backbone	黄斑区平均交并比	黄斑区中心距离差(像素)	黄斑区 AP
resnet18	0.73015±0.12564	22.81798±46.58241	0.983
resnet34	0.72774±0.12550	21.32242±28.68301	0.991
resnet50	0.72468±0.12584	21.20198±22.96725	0.989
resnext50_32x4d	0.72767±0.12232	22.54894±60.75848	0.978

resnet101	0.72453±0.12991	21.31733±30.59423	0.987
resnext101_32x8d	0.73141±0.12747	21.70120±31.41364	0.987

表 4-4 不同骨干下的视盘精度

Backbone	视盘平均交并比	视盘中心距离差(像素)	视盘 AP
resnet18	0.87607±0.05631	8.84816±5.67925	0.990
resnet34	0.87845±0.05752	8.57125±5.00088	0.988
resnet50	0.87735±0.05645	8.54313±5.19287	0.990
resnext50_32x4d	0.87930±0.05677	8.67726±4.96277	0.990
resnet101	0.87700±0.05953	8.52610±5.15709	0.988
resnext101_32x8d	0.88016±0.05421	8.35428±4.87998	0.990

由表 4-3、4-4 可以看到，黄斑区定位效果相较视盘更加不稳定，平均交并比和中心距离差的标准差较大。一是反映了黄斑区结构特征的不确定性，难以确认黄斑区的具体边界；二是由于数据集存在噪声问题，影响了模型对黄斑的分类和定位。此外，不同骨干的预训练特征提取网络之间模型性能的差距较小，这意味着可以使用较为轻量的网络来加快整个模型的运行速度。

第五章 模型部署和改进方向

5.1 简单可视化实现

本章中，本文将利用第四章中实验效果最佳的模型进行最后的项目可视化。根据模型输出结果，Imgaug 库将被用来进行目标框的生成；Flask 框架将被用来进行模型的 web 部署。

5.1.1 目标边界框

输入模型的图片数据，经正向传播后将会返回目标分类类别、分类置信度和边界框四个端点坐标的列表。将选出黄斑区和视盘分类置信度最高的边界框作为最终结果，若列表中不含黄斑区或视盘，则认为其不存在。

Imgaug 是一个封装的进行图像增强的 python 库，支持关键点（keypoint）和边界框一起变换的功能。由于原数据集中已经含有大量噪声，不再使用 Imgaug 进行数据增广，仅使用其中的目标检测框绘制函数，用于结果边界框的绘制。绘制的结果如图 5-1 所示。

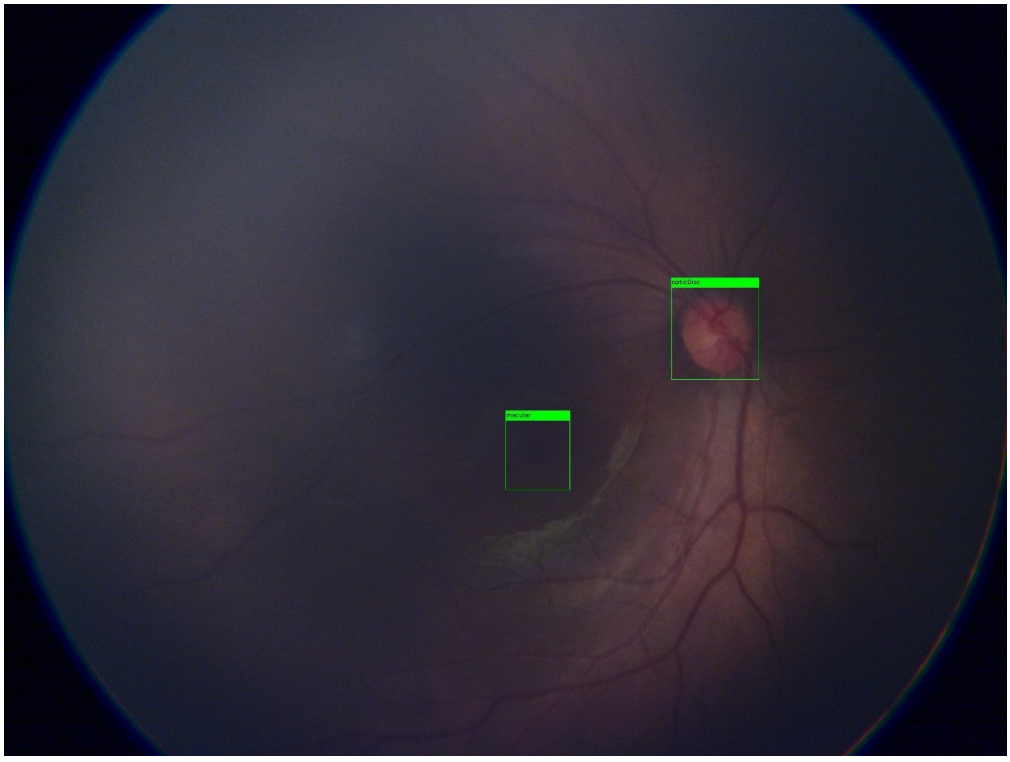


图 5-1 目标检测框的绘制　macular 为黄斑区，opticDisc 为视盘。

5.1.2 黄斑区及视盘检测的 web 部署

本文将模型构建、传播和边界框绘制整合成了一个接口，并加入了根据检测阈值输出结果的功能。4.1.1 节谈到输出结果使用分类置信度最大的边界框，而实际上因为图像质量和病变影响，结果中常含有置信度低于 0.5 甚至低于 0.1 的情况。因此，当需要高精度检测时，可以调高阈值，选择输出置信度较高的结果；当图像肉眼都难以识别且需要推测结果时，可以降低阈值，输出黄斑区或视盘可能所在的位置和大致边界，为人工识别提供帮助。该接口将配合 Flask 框架进行项目简单可视化实现。

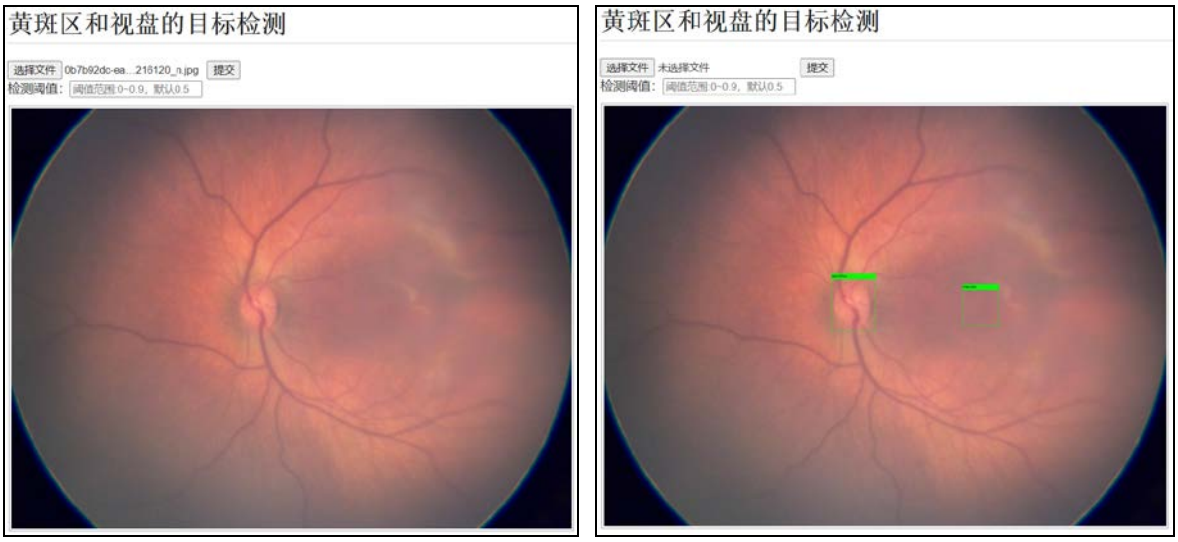
Flask 是一个使用 Python 编写的轻量级 Web 应用框架，模板引擎则使用 Jinja2。编写了一个较为简单的 html 界面（图 5-2）结合 Flask 路由机制进行黄斑区及视盘检测的 web 部署。

图 5-2 可视化操作界面

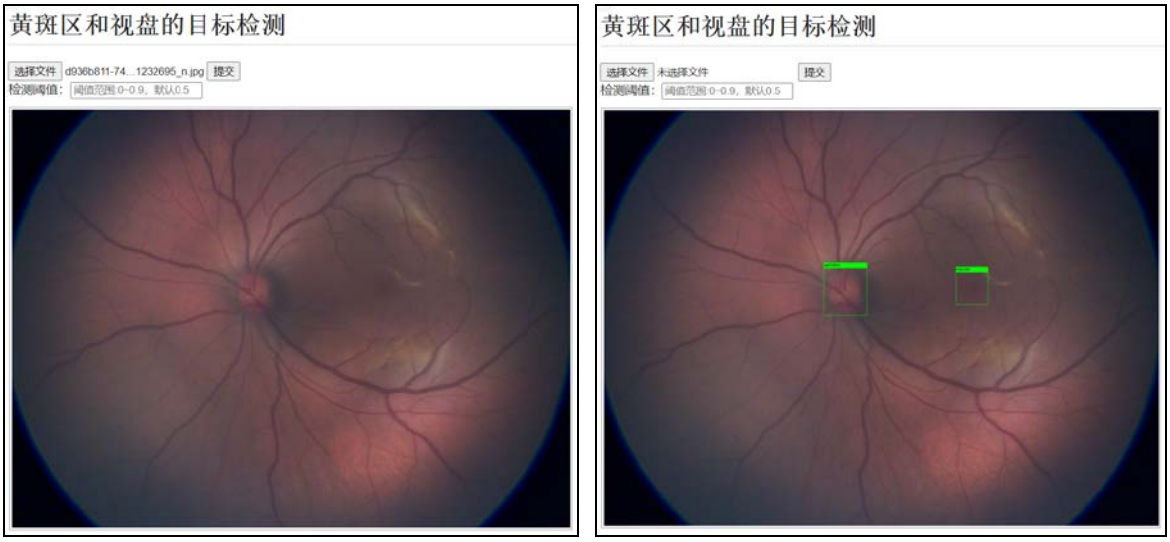
视网膜眼底图像数据由前端传至后端时，会被暂时保存在服务器本地，方便进行后续操作。接着将进行模型的构建，并将图片输入到最优模型中进行正向传播，得到的结果将

被绘制在原图上。通过把结果转换为 base64 编码返回到前端，在页面模板上进行渲染后，完成检测。

接下来对两组较为清晰的眼底图像进行功能测试：



(a)第一测试组

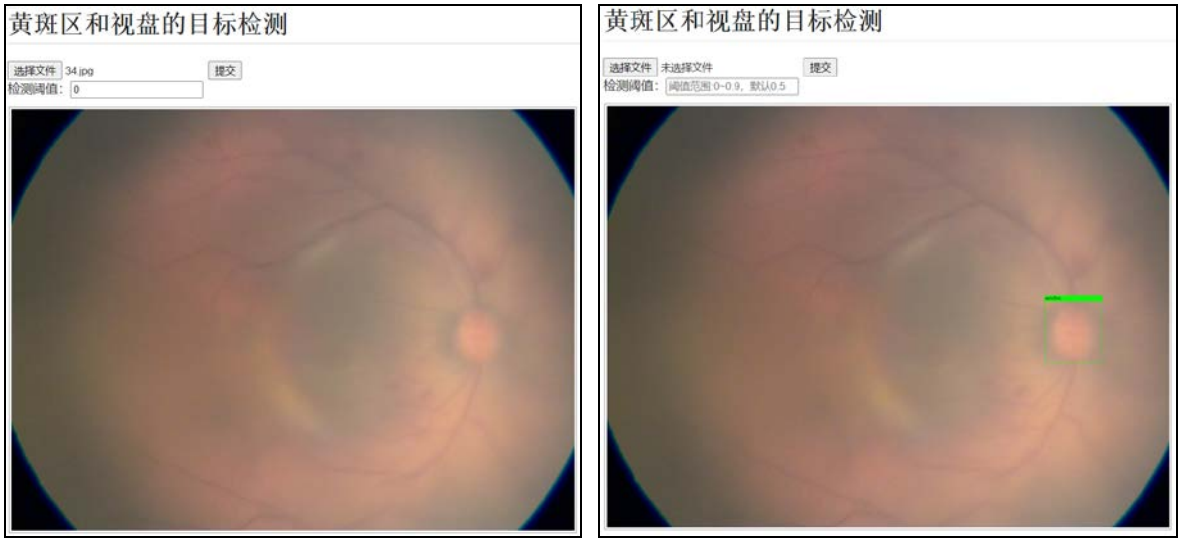


(b)第二测试组

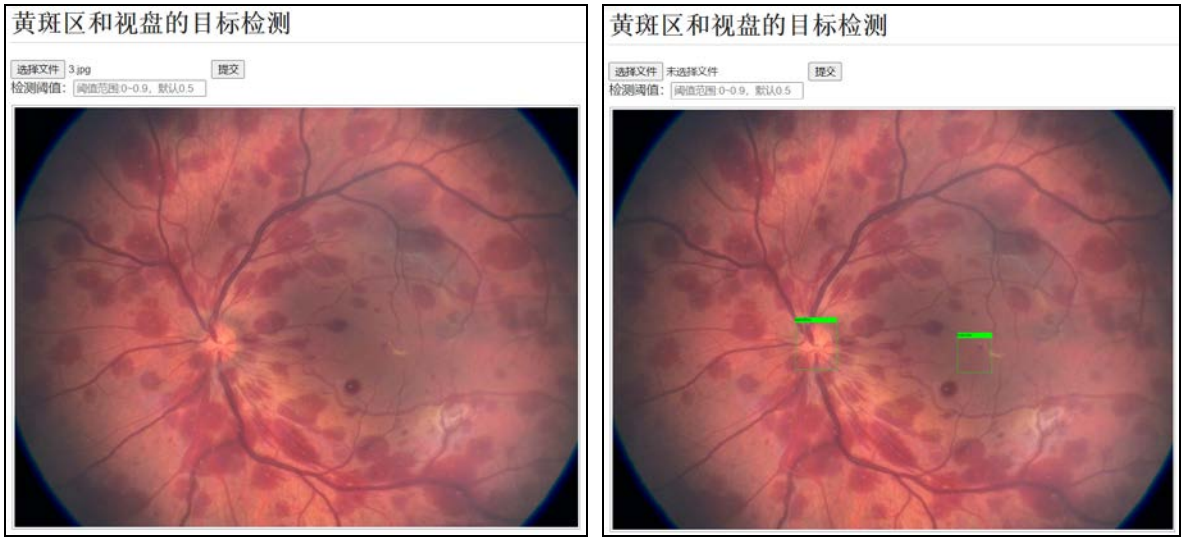
图 5-3 功能测试　左图为上传原图数据和预设检测阈值，右图为输出结果。

可见对于较为清晰且无病变的图像，即使是光线强度不同，本文的模型也能够精确地进行目标的分类和定位。然而实际上的早产儿视网膜眼底图像质量并没有如此理想，其中

存在大量因拍摄引起的图像模糊，或是病变者的图像上存在着大量出血、色斑等病变。为了验证实际应用中该目标检测系统的可行性，接下来将对较为模糊和带有病变的图像进行测试比较。

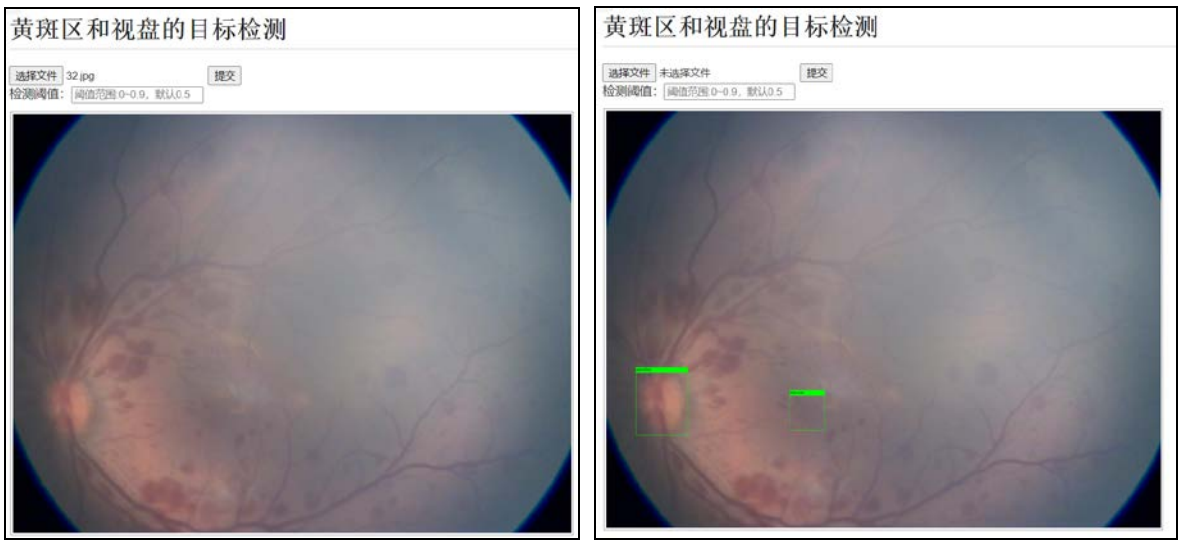


(a)极端模糊



(b)出血病变

图 5-4 性能测试　左图为上传原图数据和预设检测阈值，右图为输出结果。



(c)模糊+病变

图 5-4 续 性能测试

在该测试中选取了三组不同的图像，分别为存在图像模糊（图 5-4-(a)），存在出血病变（图 5-4-(b)）和同时存在病变和模糊问题（图 5-4-(c)）的视网膜眼底图像。第一组中可以发现在图像质量极低的情况下，黄斑区的结构特征几乎不复存在且周围的血管也模糊得难以识别，即使将阈值调至最低也无法检测到图像中的黄斑区，而视盘因其结构和轮廓较为清晰，不受图像质量影响依旧能够被检测；第二组中的出血病变对检测结果几乎无影响，可以进行精确的定位；第三组使用的较为模糊且带有出血病变的眼底图像中，虽然黄斑区的特征不如第二组中清晰，周围的血管较为模糊并存在些许出血病变，但模型依然可以精确地识别出黄斑区的所在区域。

综上所述，该检测系统能够在绝大多数的情况下进行早产儿视网膜图像的黄斑区及视盘检测和精确定位。但是当图像质量过于低下时，会出现无法识别黄斑区的情况。

5.2 未来改进方向

本文提出的改进 Mask R-CNN 网络模型已能够在实际数据集上进行黄斑区及视盘的精确检测与定位，并投入实际应用中。但是本毕业设计依然存在许多改进的地方：

距离损失函数的优化：

目前使用的距离损失函数存在着许多问题，一是基于置信区间的距离约束在模型处于高精度时效果甚微，同时 95%置信区间的设定过于直觉；二是指数函数给模型带来的难以收敛和不稳定性。后续可以考虑对置信区间进行消融实验来选择最优参数，并将损失函数换成其他较为平滑的函数如对数函数等。此外可以结合骨干网络中提取的其他特征信息进行距离约束，而不是单单局限于实际的测量数据结果。

网络结构的优化：

目前的网络结构过于复杂，特别是双头检测头中的多层卷积模块，十分影响模型训练和运行的速度。可以将原卷积层改进为轻量型卷积，譬如 MobileNet 中的 Depthwise(DW)卷积与 Pointwise(PW)卷积，在性能大致相同的情况下减少参数的数量，加快检测速度。

结合血管分割的黄斑区定位：

在图像质量较差的情况下黄斑特征难以提取，如何进行黄斑区的检测是一个重要问题。目前看来，端对端算法较难以解决这一问题。于是考虑到黄斑区周围血管密集这一特点，可以尝试结合血管分割的结果，取血管较为密集的区域作为备选候选区域加入到算法中，理论上可以提高黄斑区的检测精度。

其他：

在条件允许的情况下可以尝试增加训练迭代次数，增加数据集的数量与质量，优化模型训练方式等。

第六章　总结

该毕业设计主要研究了关于早产儿视网膜眼底图像的黄斑区及视盘检测，通过改进头部结构和加入距离监督基于 Mask R-CNN 设计了一个改良模型，并使用 Flask 简单部署到 web 端。本文提出的网络模型相较传统的图像处理算法，无需人工进行提取特征，而且精度更高检测速度更快，在实际早产儿眼底图像数据集上也显示出了极高的精度。不过，该设计也存在许多不足之处和提升空间，如改进距离损失函数进一步提高模型精度、使用轻量卷积神经网络简化网络结构等。希望该研究在未来能够对早产儿视网膜病变的诊断与治疗有所帮助。

参考文献

[1]黄旭东. 视网膜眼底彩照中视盘与黄斑定位方法研究[D].苏州大学,2019.DOI:10.27351/d.cnki.gszhu.201

[2]蒋芸, 彭婷婷, 谭宁等. 基于 YOLO 算法的眼底图像视盘定位方法*[J]. 计算机工程与科学, 2019, 第 41 卷(9):1662-1670.

[3]柯溢. 基于深度学习的视盘定位与分割研究[D].武汉轻工大学,2021.DOI:10.27776/d.cnki.gwhgy.2021.000218.

[4]汤一平, 王丽冉, 何霞等. 基于区域建议策略的视盘定位方法[J]. 中国生物医学工程学报, 2019, 第 38 卷(1):9-17.

[5]万程,周雪婷,周鹏,沈建新,俞秋丽.基于深度学习的眼底图像视盘定位与分割方法[J].中华眼底病杂志,2020,36(8):628-632.

[6]王文吉. 早产儿视网膜病变[D]. , 1996.

[7]王宪保, 朱啸咏, 姚明海. 基于改进 Faster RCNN 的目标检测方法[J]. 高技术通讯, 2021, 第 31 卷(5):489-499.

[8]杨帆, 陈睿诗, 莫阳等. 基于深度学习的视网膜病变眼底图视盘自动定位与分割研究*[J]. 贵州医科大学学报, 2020, 第 45 卷(4):432-437.

[9]张贵英,张先杰.基于深度学习的视盘自动检测[J].贵州师范学院学报,2017,33(03):27-32.DOI:10.13391/j.cnki.issn.1674-7798.2017.03.007.

[10]郑绍华, 陈健, 潘林等. 眼底图像中黄斑中心与视盘自动检测新方法[J]. 电子与信息学报, 2014, (11):2586-2592.

[11] He K, Gkioxari G, Dollar P, et al. Mask R-CNN, international conference on computer vision[J]. 2017.

[12] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28: 91-99.

[13] Wu Y, Chen Y, Yuan L, et al. Rethinking classification and localization for object detection[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10186-10195.

[14] Qureshi R J, Kovacs L, Harangi B, et al. Combining algorithms for automatic detection of optic disc and macula in fundus images[J]. Computer Vision and Image Understanding, 2012, 116(1): 138-145.

[15] Aquino A, Gegundez M E, Marin D. Automated optic disc detection in retinal images of patients with diabetic retinopathy and risk of macular edema[J]. International Journal of Biological and Life Sciences, 2012, 8(2): 87-92.

[16] Welfer D, Scharcanski J, Marinho D R. A morphologic two-stage approach for automated optic disk detection in color eye fundus images[J]. Pattern Recognition Letters, 2013, 34(5): 476-485.

[17] Lu S, Lim J H. Automatic optic disc detection from retinal images by a line operator[J]. IEEE Transactions on Biomedical Engineering, 2010, 58(1): 88-94.

[18] Gagnon L, Lalonde M, Beaulieu M, et al. Procedure to detect anatomical structures in optical fundus images[C]//Medical imaging 2001: Image processing. SPIE, 2001, 4322: 1218-1225.

[19] Köse C, İkibaş C. Statistical techniques for detection of optic disc and macula and parameters measurement in retinal fundus images[J]. Journal of Medical and Biological Engineering, 2011, 31(6): 395-404.

[20] Godse D A, Bormane D S. Automated localization of optic disc in retinal images[J]. International Journal of Advanced computer science and Applications, 2013, 4(2).

[21] Sekhar S, Al-Nuaimy W, Nandi A K. Automated localisation of optic disk and fovea in retinal fundus images[C]//2008 16th European Signal Processing Conference. IEEE, 2008: 1-5.

[22] Walter T, Klein J C, Massin P, et al. A contribution of image processing to the diagnosis of diabetic retinopathy-detection of exudates in color fundus images of the human retina[J]. IEEE transactions on medical imaging, 2002, 21(10): 1236-1243.

[23] Giancardo L, Meriaudeau F, Karnowski T P, et al. Exudate-based diabetic macular edema detection in fundus images using publicly available datasets[J]. Medical image analysis, 2012, 16(1): 216-226.

致谢

本文是在吴梦麟老师的悉心指导下完成的。从论文的选题到项目的实现，再到论文的撰写，吴老师给了我许多帮助和建议，譬如数据的清洗方法、网络的创新点、模型的改进思路、论文的写法等等。而且为了正常进行模型的训练，吴老师为我提供了实验室的高算力显卡的使用权限。同时也要感谢沈鼎学长，在实验上也给予了我许多指导。十分感谢以上两位的辛勤付出让我顺利地完成此次毕业设计和论文的撰写。