摘要

随着计算机视觉的发展，实例分割任务的应用场景与日俱增，因此实例分割技术受到研究者们的关注与研究。但目前为止该任务依然没有达到令人满意的效果，其中大部分算法在面对复杂环境时准确度不够高，与人眼识别分割还有一定差距。为解决在工程中无法使用具有高精度的实例分割算法，本论文在 Mask R-CNN 算法的基础上进行了改进，目的是解决实例分割技术在面对多目标遮挡、低照度图像、小目标等情况时能够有效进行实例分割，进一步提高算法准确度。本文的主要研究内容和贡献如下：

（1）本文针对多目标重叠问题，以 Mask R-CNN 算法为基础进行改进，设计了一种基于 Mask-RCNN 的多尺度双层分解模型（MBDN）。该算法采用 Mask R-CNN 检测头作为本文算法检测的基础部分，分割部分将重叠的对象解耦成两个图像层，其中顶层处理遮挡对象，底层处理目标对象，在每一层中增加了多尺度膨胀卷积和图卷积来进行处理，能够有效获取图像更多特征，分别从全局和局部的视角对图像中感兴趣的特征信息进行编码解码，提升算法获取全局和局部特征的能力，进而提高图像的分割性能，有效解决多目标遮挡物实例分割的问题。在主干网络 ResNet50 和 ResNet101 上，该模型分别对基线模型的精度提高了 7.3%和 7.2%，方法在 COCO 公开数据集实验中得到验证。实验结果表明，本文所提出的方法优于同类现有方法。

（2）针对低照度图像实例分割的问题,本文在 MBDN 模型基础上设计了一种基于低照度图像增强的实例分割方法。该方法在检测时采用图像增强模块，使实例分割图像亮度提升，增加其可检测性，同时将 CBAM 注意力模块加入至检测头中，提高模型对小目标的检测能力。该模型实验结果优于同类现有方法，AP 从 36.85 提升到 39.33，提高了 6.7%，有效解决复杂环境下图像实例分割问题。

综上，本文针对复杂环境下图像实例分割问题进行了探索与研究，提出了两个基于深度学习的实例分割模型，与其他方法相比效果良好，具有一定的借鉴意义。

关键词：实例分割；深度学习；多尺度卷积；重叠目标；

Abstract

With the development of computer vision, the application scenarios of the instance segmentation task are becoming more and more frequent, so the instance segmentation technique has received much attention and research from researchers. However, the task has not yet achieved satisfactory results, and most of the algorithms are not accurate enough in the face of complex environments and still fall short of human eye recognition segmentation. In order to solve the problem of using instance segmentation algorithms with high accuracy in engineering, this thesis improves on the Mask R-CNN algorithm, with the aim of solving the problem that instance segmentation techniques can effectively perform instance segmentation in the face of multiple target occlusions, low illumination images, small targets and other situations, and further improve the accuracy of the algorithm. The main research contents and contributions of this paper are as follows:

(1) In this paper, Mask-RCNN based multi-scale double layer decomposition model (MBDN) is designed for the multi-target overlap problem, based on the Mask R-CNN algorithm for improvement. The algorithm adopts the detection head of Mask R-CNN as the basic part of the detection of this paper's algorithm, and the segmentation part decouples the overlapping objects into two image layers, where the top layer deals with the occluded objects and the bottom layer deals with the target objects, and adds multi-scale expansion convolution and graph convolution in each layer for processing, which can effectively acquire more features of the image and decode the image from the global and local perspectives of interest respectively The algorithm is enhanced to obtain global and local features, thus improving the segmentation performance of the image and effectively solving the problem of segmenting multi-target occlusion instances. On the backbone networks ResNet50 and ResNet101, the model improves the accuracy of the baseline model by 7.3% and 7.2% respectively, and the proposed method is validated in COCO public dataset experiments. The experimental results show that the proposed method outperforms similar existing methods.

(2) To address the problem of low-illumination image instance segmentation, this paper designs an instance segmentation method based on the MBDN model for low-illumination image enhancement. The method uses an image enhancement module in detection to increase the brightness of the instance segmentation image and increase its detectability. At the same time, the CBAM attention module is added to the detection head to improve the detection capability of the model for small targets. The experimental results of the model

outperformed similar existing methods, with the AP improving from 36.85 to 39.33, an improvement of 6.7%, effectively solving the problem of image instance segmentation in complex environments.

In summary, this paper has explored and researched the image instance segmentation problem in complex environments, and proposed and validated two deep learning-based instance segmentation models, which are effective compared with other methods and have certain reference significance.

Keywords: instance segmentation; deep learning; multiscale convolution; overlapping targets;

目录

第一章　绪论

§1.1 研究背景与意义

视觉是人类感知世界的重要媒介，人类可以通过视觉区分不同类别的物体,而图像是利用眼睛看到的结果，是视觉感知世界的产物。通过视觉能快速接收环境信息，并对大脑内接收的图像信息加以处理，形成环境认知。因此，视觉在人类与环境交互过程中起着关键作用。随着信息技术的发展，人类正在进入信息时代，计算机将越来越广泛地进入几乎所有领域[1]。研究者们开始思考如何利用计算机设备来代替人类感知世界，其中最关键技术是计算机设备如何理解通过图像传感器获取图像信息，因此产生了计算机视觉。从字面上理解，计算机视觉是计算机拥有视觉感知的能力，衍生来说，是通过使用计算机及摄像头、相机等相关设备对图像进行处理，处理后的图像能更适合人眼观察或仪器检测，从而实现对人类视觉进行模拟。其目的在于让计算机通过摄像设备实现像人一样用视觉来感知世界，并通过计算机分析能够理解世界，能够自主的适应周围环境。计算机视觉的出现加速了人工智能的发展，使之进入感知智能时代[2]。

计算机视觉涵盖图像处理、目标检测、图像分割等领域。图像识别是指通过计算机将图像处理、理解和分析，以识别不同模式下的目标和对象[3]。然而图像识别只能对识别的目标和对象进行分类，无法定位多个对象并识别每个实例具体位置，目标检测技术可以解决这一问题。目标检测技术结合了目标定位和识别检测两项技术，在给定图像中实现目标边框的精准定位并检测出该目标所属的具体类别。目标检测技术的出现，给计算机视觉的发展带来了新的发展空间，但目标检测在对图像进行分析时，缺少对图像语意的理解，例如识别图像中马路与人群，识别图像中每个行人等。图像分割在对图像语义理解方面却有很大的进展。它可通过对组成图像的结构、颜色、灰度以及纹理等特征进行分析，将图像分成多个区域，并对于单个区域来说，这些特征具有相似性。对于目标检测而言，图像分割对图像识别任务更为精细，可以解决更加精准的目标定位、以及深入的图像语义理解等任务。

图像分割是根据像素进行分类的问题，具体是将图片中特定的类别像素分类的过程。图像分割技术在现阶段主要包括两个领域，一是语义分割，二是实例分割。实例分割技术在进行像素级别的分类的同时，还需在具体的类别基础上区别开不同的实例，是一种比较复杂技术，然而其在自动驾驶、工业机器人等领域上起着非常重要的作用。而语义分割是指仅考虑像素类别不分割同一类的不同实体[4]。在自动驾驶技术方面，实例分割系统是自动驾驶技术的核心算法之一。输入车载摄像头或者激光雷达获取的

图像到神经网络模型中，计算机通过实例分割技术将图像进行分割并归类，以实现车辆对行人和其他车辆等障碍物进行避让[5]。此外，实例分割标注任务主要是针对像素级水平上，把图像中的目标按类分别标注。这些类别可能包括行动物、人、植物等。例如，对于自动驾驶车辆来说，实例分割可以识别其在图片中的可行驶区域。在工业机器人方面，尤其在自动化装配工作中，使用实例分割技术能够在不同背景下检测和识别出零件。它可以提高整个制作过程的效率，可以很好的降低劳动成本[6]。在地理信息系统中，实例分割技术可以帮助设备在卫星遥感影像中识别道路、楼房、草地等类别，并对识别的对象在图像中标注每个像素。例如，可以检测地图上生态系统变化。在卫星地图导航方面，实例分割也起到了很大的作用。在医疗方面，由于人工智能的不断发展，智能医疗研究逐渐成熟，将图像处理与医疗影像分析诊断相融合也成为研究热门[7]。将新技术应用到这类领域，可以帮助医生更好的工作[8]。基于此背景，本文研究实例分割技术对于智能工业发展具有重大意义，实现智能感知，提高人们工作以及生活效率。

§1.2 国内外研究现状及发展趋势

近年来，目标检测、语义分割成为计算机视觉任务的研究热点，而将目标检测与语义分割相结合就成为实例分割。目标检测主要任务是检测并框选出实例目标，语义分割主要任务是对每个像素进行分类，得到不同所属类，实例分割通过分割可数目标对象并提供语义和实例标签，从而区分相同语义类别中不同实例。语义分割和目标检测是实例分割形成的基础，因此，本文将从目标检测、语义分割、实例分割三个方面来阐述实例分割的研究现状及发展趋势。

§1.2.1 目标检测

目标检测作为计算机视觉的一项任务，早在 2005 年 HOG 检测器的提出就已经出现。方向梯度直方图（Histogram of Oriented Gradient，HOG）[9]对当时主流的尺度不变特征变换、边缘方向直方图和形状上下文进行改进，通过计算图像梯度方向信息的统计值实现对目标特征描述。尽管在检测各种对象类时 HOG 已经被广泛使用，但为检测不同大小对象，在检测窗口保持大小不变的同时，HOG 检测器还需多次对输入图像重新进行标度。手工特征需要大量人工干预，且性能趋于饱和，目标检测发展缓慢，检测能力没有较大提升。

直至 2012 年，卷积神经网络（Convolutional Neural Network，CNN）[10]的出现，在全球学术界掀起了轩然大波，利用 CNN 实现人工智能各项任务，成为学者们研究的重点。由于深度卷积网络具有强大的特征提取能力，可以有效学习图像的鲁棒性与

高层次特征表示，因此开始将深度学习算法应用于目标检测。于是在 2014 年，R. Girshick 等人[11]首先提出了目标检测中使用拥有卷积神经网络特征区域的 RCNN，成为将深度学习成功应用于目标检测上的第一个算法，至此目标检测开始迅速发展。尽管 RCNN 取得巨大进步，但它在冗余的特征计算中存在大量重叠，致使检测速度极其缓慢，因此，为解决这一难题，2014 年，K. He 等人[12]提出了空间金字塔池化网络（Spatial Pyramid Pooling Networks，SPPNet）。SPPNet 通过引入了空间金字塔池化（Spatial Pyramid Pooling，SPP）层，使卷积神经网络能够生成固定长度的表示，其速度是 R-CNN 的 20 多倍，检测精度稳定且不需要重新缩放图像和感兴趣区域的大小，避免了卷积特征重复计算， 但 SPPNet 只对全连接层之前的层进行了微调，忽略了其作用。R. Girshick 等人[13] 2015 年提出了 Fast RCNN，对 R-CNN 和 SPPNet 进一步进行改进。在同样的网络参数中，它可以对检测器和边界框回归器进行同时训练。尽管 Fast-RCNN 成功将 R-CNN 和 SPPNet 的优点融合，但检测速度依然受限。于是 Ş. Ren 等人[14]提出了 Faster RCNN 检测器，它是第一个端到端的，且接近实时的深度学习检测器。虽然 Faster RCNN 后续的检测阶段仍然存在计算冗余，但从 RCNN 到 Faster RCNN，目标检测系统中大部分独立模块都已经逐渐整合成统一的端到端学习框架。然而在深度学习检测器中，大部分检测器还只在网络顶层进行检测，不利于对象的定位。针对此问题，T.-Y.Lin 等人[15]在 Faster RCNN 基础上提出了特征金字塔网络。FPN 拥有横向连接的自上而下体系结构，卷积神经网络通过正向传播，形成特征金字塔，在各种尺度目标方面的检测有很大提高，目前成为了许多最新探测器的不可分割的一部分。第一个单阶段检测算法是 R. Joseph 等人[16]提出的 YOLO。它将单个神经网络应用于整个图像，将图像分割成多个区域，同时对每个区域的边界框和概率进行预测，因此其检测速度非常快。相比于两级探测器，它的探测速度有了很大的提升，但定位精度有所下降。W. Liu 等人[17]提出 SSD，主要是增加了多参考和多分辨率检测技术，对于单级检测器检测精度有很大的提升，尤其对于部分小目标，在检测速度和准确度上都有优势。尽管由于二级检测器的精度高，常年优于单机检测器，但单级检测器速度快、结构简单。

§1.2.2 语义分割

语义分割一直是图像理解领域的热点，但也是计算机视觉研究中的经典难题[18]。语义分割是由图像分类、目标检测和图像分割结合而成，使用一定方法通过将图像分割成具有一定语义含义的区域块，针对每个区域块识别出语义类别，完成从底层到高层的语义推理过程，最终得到一幅具有逐像素语义标注的分割图像。在图像语义分割方法中，主要包含两种方法，分为传统方法和基于卷积神经网络方法。在传统分割算法中，主要包含灰度分割，条件随机场等算法。然而深度学习的出现大幅度提高了分

割精度。FCN 作为一个里程碑式图像分割网络，它证明了可以通过端到端的方式在可变大小的图像上训练深层网络进行语义分割。但由于 FCN 感受野固定和分割物体细节时容易出现丢失或被平滑问题，导致忽略全局。针对忽略全局的问题，Liu 等人[19]提出了 ParseNet 模型，模型解决了此问题，主要利用使用层的平均特征来增加每个位置的特征，从而将全局添加上下文到 FCN。目前，许多分割问题都在使用 FCNs 如脑肿瘤分割、皮肤损伤分割和虹膜分割等。Cambridge 等人[20]提出了 SegNet。SegNet 网络解决了 FCN 在语义分割时感受野固定和分割物体细节时容易出现丢失或被平滑的问题，保存图像轮廓信息采用了 pooling indices，降低了参数数量。但由于 maxpool 过于采用直接删除特征值方法，使得信息后续难以恢复。Ronnebergeretal[21]提出了 U-Net，主要用于分割生物显微镜图像，其网络和训练策略是通过使用数据增强来更有效地从可用的注释图像中进行学习。由于在卷积过程中没有加填充，特征长度在每一次卷积后就会减少两个像素，致使最后输入与输出大小不一致。Chen 等人[23]提出 Deep Lab V1 网络，该网络使用空洞卷积来代替深度卷积神经网络（DCNN）的部分卷积层，在不增加参数的情况下增大感受野，获得了更多特征信息。此外，为增强获取图像细节信息的能力，实现目标精确定位，在 DCNN 最后一层添加了全连接条件随机场。Chen 等人[23]对 Deep Lab V1 进行扩充提出了 Deep Lab V2，通过结合空洞卷积和空间金字塔池化模型，提出了带孔空间金字塔池化（ASPP）模块。ASPP 模块为获取多种尺度的特征，采用多种不同采样率的空洞卷积，并进行特征融合从而获得上下文信息，实现多尺度目标的处理。DeepLabv3 语义分割模型与 Deep Lab V2 相比，在 ASPP 中增加了全局平均池化，与此同时，将批量归一化在平行扩张卷积后添加，更好地捕获了全局语境信息。DeepLabv3+是在 DeepLabv3 的基础上进行改进，增加了编解码模块和 Xception 主干网络，增加编解码模块目的在于更好地恢复原始的像素信息，保留分割的细节信息，同时丰富了编码上下文信息。RefineNet[24]通过详细划分中间激活映射并分层地将其连接到多尺度激活，防止锐度损失。最近，半监督和弱监督的语义分割技术越发成熟，人们正在尝试使用不可靠标签和弱标注的情况下进行语义分割，来减少模型训练的工作量，使其更适合在工业上的应用。例如，Lixiang Ru 等人[25]提出了一种使用 Transformer 的端到端弱监督语义分割方法，该方法引入视觉 Transformer 结构，并探索了适合视觉 Transformer 的初始伪标签生成方法。但这些方法与全监督的方法相比在分割精度上有些差距，特别是在低照度、有遮挡物及小目标的情况下，识别效果还需加强。

§1.2.3 实例分割

与其他视觉经典任务相比，实例分割（Instance Segmentation）[26]是最难的一个，它不但包含语义分割的特点，需要像素层面上进行分类，而且含有目标检测中部分特

点，尽管是同一种类，也需要定位出不同实例。因此，在两阶段方法中，针对于方向的不同，实例分割的研究分为两条线，分别是自上而下在检测的基础上进行研究和自下而上的在语义分割的基础上进行研究。自上而下方法是使用目标检测方法预测顶层边界框，在图像中通过目标检测器确定实例的位置，同时目标检测器也可以确定每个实例属于哪个类别，对每个边界框内进行分割，结果作为实例掩膜进行输出，因此，目标检测在此类方法中尤为重要。自下而上方法通过将每个像素映射为向量，再使用聚类方法将向量嵌入到不同的实例中。此方法需要高质量的像素级映射结果和精心设计的聚类方法，并且后处理方法泛化能力差，不能处理复杂的图像情况，特别是对象被遮挡或者重叠的情况。

最早的实例分割算法是由 Hariharan B[27]提出的 SDS，通过推荐生成、特征提取、区域分类和区域改良来获得分割结果，但具有很大的局限性，只用卷积神经网络技术进行特征提取的弊端在于提取的掩码细节粗糙，位置信息不精确。虽然分割效果不理想，但是 SDS 提出的实例分割方法为后面实例分割研究有了很大的启发。在 2017 年，He K 团队[28]基于 Faster RCNN 分类与回归分支又添加一个分支，目的是用于语义分割，进而提出了 Mask RCNN 算法。与 Faster RCNN 相比，Mask RCNN 使用优秀的 ResNet-FPN 结构作为基础网络结构，多层特征图对于多尺度物体以及小物体的检测更有利，Mask RCNN 算法提出 Rol Align 方法来代替 Rol Pooling，取消了取整操作，保留了浮点数，并基于分类和回归添加一个掩码分支从而预测每一个像素类别，采用全卷积网络(Fully Convolutional Network，FCN)[29]网络结构，通过卷积与反卷积建立端到端网络，并对每一个像素分类，从而实现了较好的分割效果。但是 Mask RCNN 也有明显的不足，过度依赖框的准确性，一些小尺度的物体检测效果较差，难以胜任边缘精度要求高的任务。Chen L C 等人[30]提出了 MaskLab，也是基于 Faster-RCNN 对象检测器进行构建，预测框提供对象实例准确定位。在每个感兴趣区域内，MaskLab 通过将语义和方向预测相结合来对前景或背景进行分割。然而，它依赖于 RPN 生成候选区域，可能会导致漏检一些小目标或密集目标。Liu S 等人[31]提出了 PANet，通过特征金字塔构建不同尺寸且有高级语义信息的特征图，同时这也增加了计算量和内存消耗。Enze Xie 等人[32]提出了一种基于密集预测的方法 PolarMask，采用极坐标表示每个对象的轮廓，利用一个回归分支预测中心点的位置和射线的距离，提高了效率和准确度。但它使用了 NMS 来删除多余的掩码，这会降低速度和准确度。Xinggang Wang 等人[33]提出一种基于弱监督的方法 BoxCaseg，采用只有边界框标注的图像和显著性图像进行联合训练，通过使用一个多任务网络生成类别无关的目标掩码，并用一种掩码合并和丢弃策略来生成伪掩码，然后用伪掩码训练一个 Mask R-CNN 模型，然而，这可能会导致任务之间的干扰和冲突，降低网络的泛化能力。Hao Chen 提出[34]了一种基于区域的方法 BlendMask，利用注意力机制将不同尺度的特征融合成一个高分辨率的张量，然后用一个解码器将张量转化为实例掩码。这样可以克服传统方法中