**目录**

第一章 绪论

1.1 研究背景及意义

1.2 本文主要工作

1.3 论文结构

第二章 图像语义分割技术研究现状

2.1 基于低级特征提取的图像分割方法

2.2 基于卷积网络的语义分割方法

2.3 基于区域选择的分割算法改进

2.4 本章小结

第三章 图像语义分割的基本算法实现

3.1 方法步骤流程

3.2 基于labelme的图像数据预处理

3.3 基于YOLACT的实时实例分割

3.3.1 核心算法与工作原理

3.3.2 数据增强代码与分析

3.3.3 训练过程参数的影响

第四章 图像语义分割的实际应用与细节优化

4.1 针对嵌套圆管的数据集制作

4.2 实验环境和参数设置

4.3 应用过程中的细节优化

4.4 效果展示与结果分析

第五章 总结与展望

5.1 工作总结

5.2 未来展望

1. **绪论**

本章探讨本文的研究背景及意义，对本文工作内容和论文结构进行简要的介绍。

* 1. **研究背景及意义**

二十一世纪是人工智能的时代，人工智能也被认为是第四次工业革命，许多全球顶尖的技术公司都把目光转向人工智能。人工智能，即模仿生物神经网络的原理建造实用的人工神经网络模型，设计相应的学习算法，模拟人脑的某些智能活动，比如视觉、听觉，以及对所视、所听的理解，甚至于思考并作出反应。借此，计算机视觉这一研究方向应运而生。

那么，人类是如何描述场景的呢？我们可能会说“窗户下有一张桌子”，或者“沙发右边有一盏灯”。图像理解的关键在于将一个整体场景分解成几个单独的实体，这也有助于我们推理目标的不同行为。

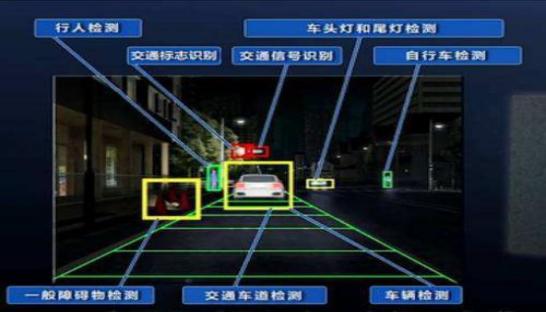
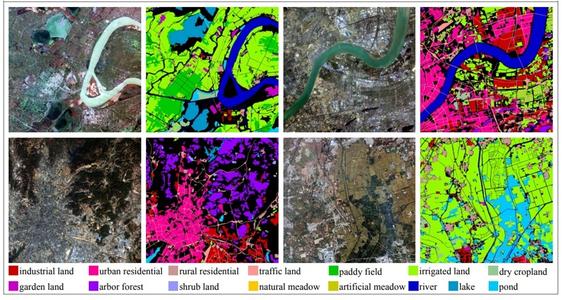
在计算机视觉领域，目前神经网络的应用主要包括图像识别、目标定位与检测以及语义分割。图像识别告诉你图像是什么，目标定位与检测告诉你图像中的目标在哪里，而语义分割则是让计算机从像素级别回答上述两个问题。虽然目标定位与检测可以帮助我们绘制某些确定实体的边框，但人类对场景的理解远远不止于此，我们能以像素级的精细程度对每一个实体进行检测并标记精确的边界，因此计算机仅仅做到识别、定位与检测是不够的，根据图像语义精确分割出实例变得越来越重要。

语义分割是计算机视觉中的基本任务，在语义分割中我们需要将视觉输入分为不同的语义可解释类别，“语义的可解释性”即分类类别在真实世界中是有意义的。例如，我们可能需要区分图像中属于汽车的所有像素，并把这些像素涂成同一种颜色，如图1-1所示。

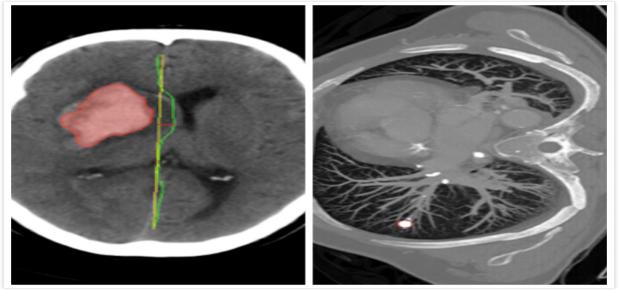


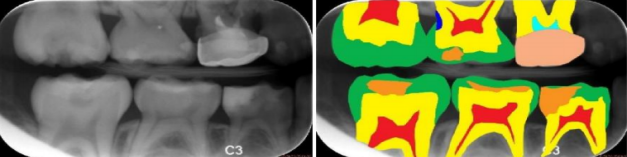
图1-1 语义分割演示视频截图

由于图像、视觉与人们生活息息相关，所以语义分割在现代生活中应用十分广泛。如图1-2（a）所示，在地理信息系统中，通过训练神经网络让机器输入卫星遥感影像，自动识别道路，河流，庄稼，建筑物等，并且对图像中每个像素进行标注；如图1-2（b）所示，在无人驾驶交通中，车载摄像头，或者激光雷达探查到图像后输入到神经网络中，后台计算机可以自动将图像分割归类，以避让行人和车辆等障碍；如图1-2（c）所示，在医疗影像分析中，通过对图像的语义分割诊断肿瘤、龋齿等不良生理现象。



（a）地理信息系统 （b）无人驾驶





（c）医疗影像分析

图1-2 语义分割在现代生活中的应用

由此可见，随着越来越多的领域以及应用程序通过从图像中推断知识来获取“营养”，“对场景的理解”逐渐成为一个核心的计算机视觉问题。而从宏观上看，在计算机视觉的应用中，语义分割相当于解决整个问题的预处理阶段，是实现场景完整理解的基础。可谓是，没有正确的分割就不可能有正确的识别。因此，语义分割是当今计算机视觉领域的关键问题之一。

语义分割是图像处理中的一项关键技术，也是一个经典难题，发展至今仍没有找到一个比较通用的方法，也没有制定出判断分割算法好坏的标准。第二章中将对近年来出现的多种语义分割方法进行较为全面的介绍，探讨语义分割技术的发展方向。

近十年以来，人工神经网络识别技术已经引起了广泛的关注，并且大量应用于语义分割领域。基于神经网络的语义分割方法的基本思想，是通过训练多层感知机来得到线性决策函数，然后用决策函数对像素进行分类来达到语义分割的目的，这种方法需要投入大量的训练数据。神经网络存在巨量的连接，容易引入空间信息，能较好地解决图像中的噪声和不均匀问题。选择何种网络结构是此类基于神经网络的技术要解决的主要问题。

* 1. **本文主要工作**

本文研究的课题是基于YOLACT的图像语义分割模型的复现以及参数的训练优化，并将模型应用于物品复杂摆放的图像上，正确地识别并标注出每一物品实例。主要工作如下：

1. 实现YOLACT的经典算法，掌握调参、训练、评估方法。
2. 根据分割目标制作自己的数据集，用于语义分割的训练集，并将模型测试的输出数据进行标签细化。
3. 将训练得到的模型进行一定量的测试，根据测试结果的不足，对模型参数进行调整，或对算法进行改进。
   1. **论文结构**

本文的各章节组织结构如下：

第一章：绪论。简要介绍了图像语义分割的研究背景和意义，并概述了本文的主要工作和行文结构。

第二章：图像语义分割技术研究现状。介绍了近年来图像识别与语义分割的发展以及国内外的研究现状，包括了模型与算法的演变过程。

第三章：图像语义分割的基本算法实现。详述了本文所重点采用的YOLACT模型的算法原理和实现思路。

第四章：图像语义分割的实际应用与细节优化。针对本文对复杂图片进行分割的任务特点，对上述模型与算法进行细节上的优化，最后展示分割效果，并分析改进后精度和效率的提升。

第五章：对本文的工作进行总结，讨论存在的问题以及对未来工作的展望。

1. **图像语义分割技术研究现状**

图像语义分割，即让计算机根据图像的语义对图像中的不同对象进行分割与标注。相关技术经过多年的研究与发展，其算法也不断迭代更新，以下将以时间顺序介绍各个时期涌现的具有代表性的算法。

**2.1 基于低级特征提取的图像分割方法**

普通的图像分割，通常意味着传统语义分割，这个时期的图像分割（大概2010年之前），由于计算机计算能力有限，早期只能处理一些灰度图，后来才能处理RGB图，这个时期的分割主要是通过提取图片的低级特征，然后进行分割，分割出来的结果并没有语义的标注，换句话说，并不知道分割出来的东西是什么。这个时期涌现出一些方法，如Ostu、FCM、分水岭、N-Cut等。这些方法给出的结果在当时还不错，但缺点也很明显，需要人工干预，这个缺点在将来批量化处理的智能时代简直就是死穴。

**2.2 基于卷积网络的语义分割方法**

之后，随着计算能力的提高，人们开始考虑分割并获得图像的语义，这里的语义目前还是低级语义，主要指分割出来的物体的类别，这个阶段（大概是2010年到2015年）人们开始引入卷积神经网络CNN，使用机器学习的方法进行语义分割。

2015年，国际计算机视觉与模式识别会议（CVPR）上发表了一篇论文，论文第一次提出了全卷积网络FCN，获得了当年的最佳荣誉提名论文。随着FCN的出现，深度学习正式进入语义分割领域，这里的语义仍主要指分割出来的物体的类别，但从分割结果已经可以清楚地知道分割出来的是什么物体，比如猫、狗等等。

对于一般的CNN网络，如VGG和ResNet，都会在网络的最后加入一些全连接层，经过softmax函数的预测后就可以获得类别概率信息。但是这个概率信息是1维的，即只能标识整个图片的类别，不能标识每个像素点的类别。

如图2-1所示，FCN则是把后面几个全连接都换成卷积，这样就可以获得一张2维的特征图，后接softmax获得每个像素点的分类信息。

1. 对于FCN-32s，直接对pool5特征图进行32倍上采样获得32倍特征图，再对32倍特征图每个点做softmax预测获得FCN-32s预测图（即分割图）。
2. 对于FCN-16s，首先对pool5特征图进行2倍上采样获得2倍特征图，再把pool4特征图和2倍特征图逐点相加，然后对相加的特征图进行16倍上采样，并做softmax预测，获得FCN-16s预测图。

对于FCN-8s，首先进行pool4 + 2x逐点相加后两倍上采样，然后又进行pool3 + 2x逐点相加，即进行更多次特征融合……

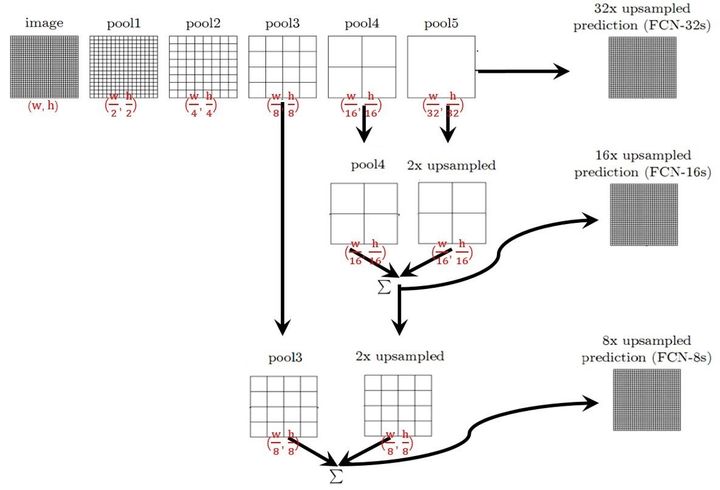


图2-1 不同倍数的FCN预测图生成过程示意图

如图2-2所示，从精度上看，FCN-32s＜FCN-16s＜FCN-8s，即使用多层特征融合有利于提高分割准确性。

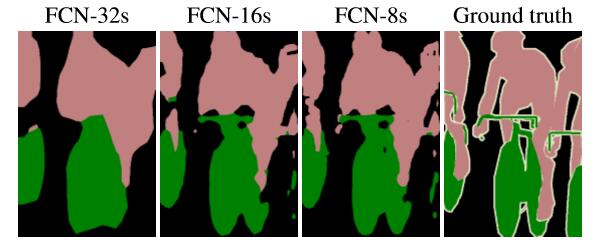


图2-2 不同倍数上采样结果对比图

因此可以看出，CNN的识别是图像级的识别，也就是从图像到结果；而FCN的识别是像素级的识别，对输入图像的每一个像素在输出上都有对应的判断标注，标明这个像素最可能是属于一个什么物体或类别。但是FCNN并不是完美无缺，主要有以下两个缺点：

1.得到的结果还是不够精细。进行8倍上采样虽然比32倍的效果好了很多，但是上采样的结果还是比较模糊和平滑，对图像中的细节不敏感。

是对各个像素进行分类，没有充分考虑像素与像素之间的关系。忽略了在通常的基于像素分类的分割方法中使用的空间规整步骤，缺乏空间一致性。

**2.3 基于区域选择的分割算法改进**

在语义分割领域，基于区域选择的几个算法主要是由前人的有关于目标检测的工作渐渐延伸到语义分割的领域的。目标检测的主要研究进程由R-CNN到Fast R-CNN再到Faster R-CNN。

Faster R-CNN是两阶段的目标检测算法，包括阶段一的Region proposal以及阶段二的Bounding box回归和分类。Faster R-CNN的整个流程如图2-3所示。

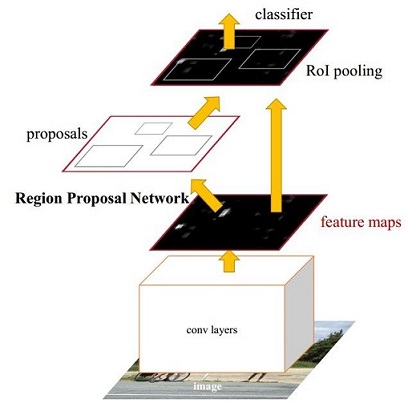


图2-3 Faster R-CNN算法流程示意图

Faster RCNN使用CNN提取图像特征，然后使用Region Proposal Network（RPN）去提取出ROI（感兴趣区域），然后使用ROI池化将这些ROI全部变成固定尺寸，再喂给全连接层进行Bounding box回归和分类预测。

在此基础上，我国何恺明团队提出一个基于Faster R-CNN模型的一种新型的分割模型Mask R-CNN，此论文斩获ICCV 2017的最佳论文。在Mask R-CNN的工作中，它主要完成了三件事情：目标检测，目标分类，语义分割。作者在Faster R-CNN的结构基础上加上了Mask预测分支，并且改良了ROI Pooling，提出了ROI Align。

**2.4 本章小结**

由本章内容可知，图像语义分割技术随着精度要求的提高而不断地改进和发展，而算法的更新迭代总是在解决前一阶段遇到的问题，这个问题可以概括为对“语义”的理解过程。从分割而不识别的基于低级特征提取的图像分割，到分割同时进行分类的基于卷积网络的语义分割，再到可以将同一类别的不同实例分割出来的实例分割，对“语义”的理解也逐渐深入透彻，逐渐考察更多的细节，逐渐向人类对图像的认知靠近。近几年不断有更新颖、更前沿的模型被提出，如Mask Scoring R-CNN、YOLACT、Blend Mask等。此次毕业设计具体选择了其中的YOLACT进行深入研究。

1. **语义分割的基本算法实现**

**3.1 方法步骤流程**

本实验所使用的方法主要包括图像数据预处理和图像语义分割训练两个阶段。一阶段是选取合适的原始数据集，使用合适的标注方法对原始数据集进行手动标签标注，并将标注后的数据文件转换为合适的格式。二阶段是将数据集投入训练，根据训练结果改进数据集的选取和标注方式。具体方法步骤流程如图3-1所示。

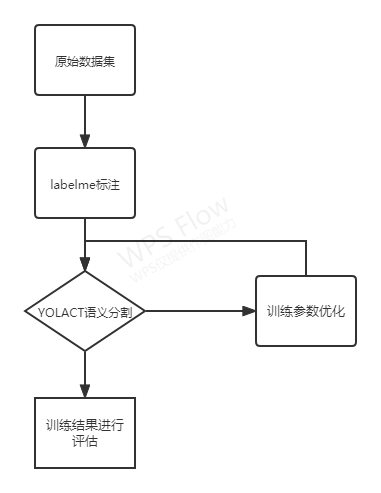


图3-1 本文实验具体方法步骤流程示意图

**3.2 基于labelme的图像数据预处理**

图像数据的预处理主要分为两个部分：

1. 使用labelme软件对数据集（包括训练集和验证集）中每张图片里的每个实例进行标记，获取对应的json文件。
2. 将第一步得到的json文件转为COCO数据集的格式。

**3.3 基于YOLACT的实时实例分割**

**3.3.1 核心算法与工作原理**

近年来的实例分割可以分为两类：一类是two-stage的方法，即“先检测后分割”的方法，首先定位到目标物体的边框，然后在边框内分割目标物体，典型的代表是Mask R-CNN；另一类是one-stage的方法，典型代表就是YOLACT。YOLACT模型并非只是在Mask R-CNN的基础上进行小修小补，而是基于one-stage全卷积算法做了重新设计，虽然在精度上稍低于Mask R-CNN，但是在精度满足大部分需求的同时，速度大大提升，达到了实时的效果，并且容易部署，可广泛应用于多种不同的场景。

YOLACT是2019年发表在ICCV上面的一个全卷积的实时实例分割模型，根据评估，该模型在处理550\*550大小的图片时，可以在29.8mAP的精度下达到33.5FPS的速度，比以往任何方法都要快得多，而互联网上多数视频一般是30FPS，这也就是“实时”的含义了。YOLACT的是You Only Look At CoefficienTs的简写，其中coefficients也就是这个模型的输出之一，这个命名风格应该是致敬了另一目标检测模型YOLO。

YOLACT模型的框架如图3-2所示，该模型非常具有创新性地摒弃了隐含的特征定位步骤，通过将实例分割分解为两个并行的子任务来实现这一点：

1. Prediction Head分支生成各个anchor的类别置信度、位置回归参数以及mask系数；
2. Protonet分支生成一组独立于实例的原型mask，然后将mask与mask系数线性结合，从而得到图片中每一个目标物体的mask。

由于这个过程不依赖于重新池化，这种方法可以产生非常高质量的mask，并以极低地代价显示出时间稳定性。

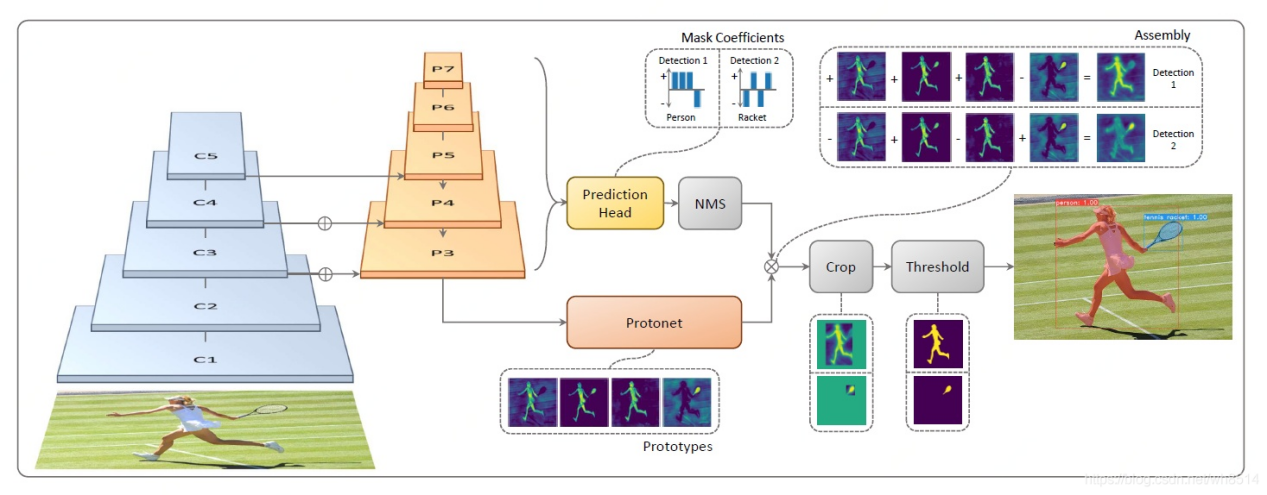


图3-2 YOLACT模型的框架示意图

1. 骨干网络ResNet101-FPN：五层卷积C1~C5，然后由C5经过1个卷积层得到P5，然后对P5采用双线性插值使特征图扩大一倍，与经过卷积的C4相加得到P4，再采用同样的方法即可得到P3。再然后，对P5进行卷积和下采样得到P6，对P6进行同样的卷积和下采样得到P7，从而建立FPN网络。P3被送入Protonet，同时P3-P7被送到Prediction Head中。
2. Protonet分支：其网络结构如图3-3所示，由若干卷积层组成，输出k个通道的prototype mask。

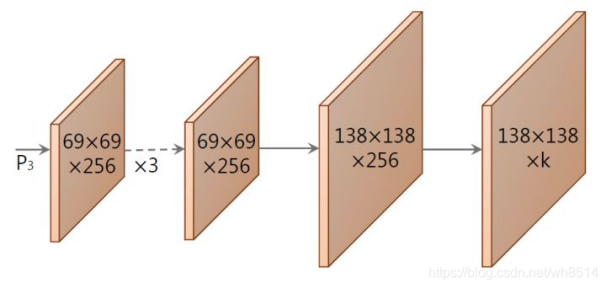


图3-3 Protonet分支的网络结构示意图

1. Prediction Head分支：其网络结构如图3-4所示，在RetinaNet的基础上改进得到的，采用共享卷积网络，从而可以提高速度，达到实时分割的目的。它的输入是 P3 ~ P7 共五个特征图，每个特征图先生成anchor，然后生成每个anchor的类别置信度、位置回归参数以及mask系数。

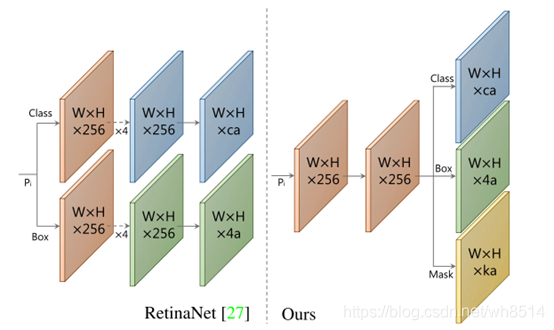


图3-4 Prediction Head分支的网络结构示意图

1. NMS：通过Prediction Head分支网络后会得到很多anchor，可以在anchor的位置加上位置偏移得到RoI位置。由于RoI存在重叠，NMS是常用的筛选算法。
2. Crop & Threshold：将通过Prediction Head分支得到的mask系数和Protonet分支得到的prototype mask做矩阵乘法，就可以得到图像中每一个目标物体的mask。Crop指的是将边界外的mask清零，Threshold指的是以0.5为阈值对生成的 mask进行图像二值化处理。

综上所述，可见YOLACT模型的算法是非常具有创新性的，大胆抛弃了Mask R-CNN那一套较为繁琐冗杂的设计，整个流程显得非常的简洁且高效，并行预测当前图片的原型掩码（prototype mask) 和每个bbox实例的掩码系数（mask coefficients)，然后通过将原型与掩模系数线性组合来生成实例掩码（instance masks)，思路很有独到之处，是近几年被提出的前沿算法中非常值得研究与学习的对象。

**3.3.2 数据增强代码与分析**

与数据增强相关的代码位于train.py文件中。首先是数据的加载，这里传入的参数包括训练集图像的路径image\_path、训练集标记文件info\_file，以及使用的数据增强方式transform。

1. dataset = COCODetection(image\_path=cfg.dataset.train\_images,
2. info\_file=cfg.dataset.train\_info,
3. transform=SSDAugmentation(MEANS))

在这里重点关注一下SSDAugmentation类中包含了哪些数据增强方法。

1. **class** SSDAugmentation(object):
2. """ Transform to be used when training. """
4. **def** \_\_init\_\_(self, mean=MEANS, std=STD):
5. #构造Compose的实例对象，传入的是一个列表
6. self.augment = Compose([
7. ConvertFromInts(), #将图片数据转为np.float32的数据类型
8. ToAbsoluteCoords(), #计算bbox的绝对坐标
9. enable\_if(cfg.augment\_photometric\_distort, PhotometricDistort()), #光度扭曲
10. enable\_if(cfg.augment\_expand, Expand(mean)), #扩张
11. enable\_if(cfg.augment\_random\_sample\_crop, RandomSampleCrop()), #随机裁剪
12. enable\_if(cfg.augment\_random\_mirror, RandomMirror()), #随机镜像
13. enable\_if(cfg.augment\_random\_flip, RandomFlip()), #随机翻转
14. enable\_if(cfg.augment\_random\_flip, RandomRot90()), #随机旋转
15. Resize(),
16. enable\_if(**not** cfg.preserve\_aspect\_ratio, Pad(cfg.max\_size, cfg.max\_size, mean)),
17. ToPercentCoords(),
18. PrepareMasks(cfg.mask\_size, cfg.use\_gt\_bboxes),
19. BackboneTransform(cfg.backbone.transform, mean, std, 'BGR')
20. ])

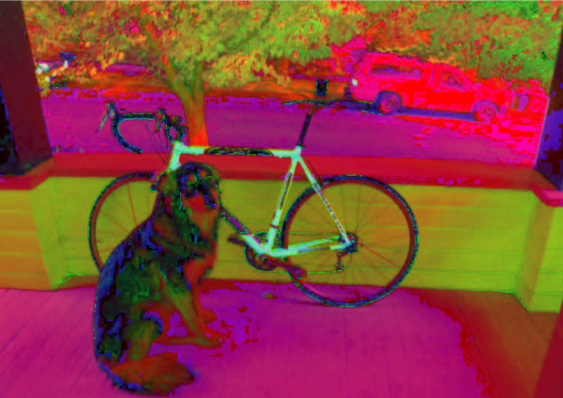
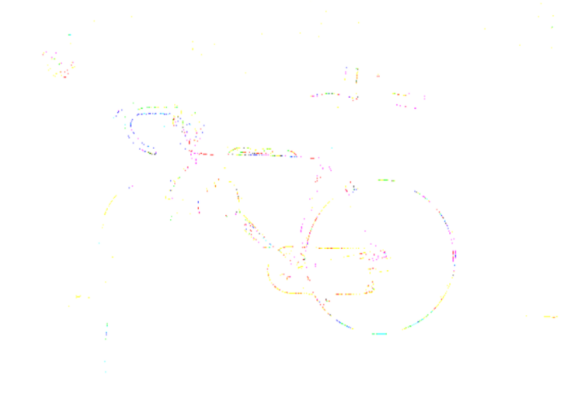
这里用到了光度扭曲、图像扩张、随机裁剪、随机镜像、随机翻转、随机旋转等数据增强方法。对图像进行几何变换的数据增强方法比较容易理解，但类似于光度扭曲的方法可能就难以想象出其效果。于是通过调用与光度扭曲相关的代码，可对这种数据增强方法进行直观地展示。光度扭曲效果如图3-5所示。



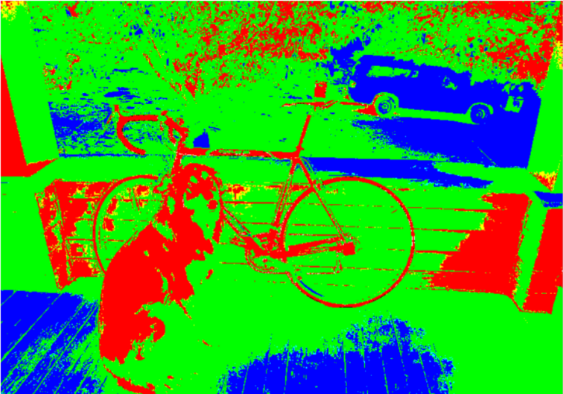
（a）处理前 （b）处理后

图3-5 光度扭曲效果图

光度扭曲实际上是由多种像素操作组合而得到的，包括随机对比度、颜色通道转换、随机亮度等操作，其效果在图3-6中分别进行展示。



（a）随机对比度处理 （b）颜色通道转换处理



（c）随机亮度处理

图3-6 三种不同的像素操作效果

数据增强在深度学习中是不可缺少的步骤，其作用有：

1. 避免过拟合。当数据集具有某种明显的特征，例如数据集中的图片基本在同一个场景中拍摄，数据增强可以避免模型学习到与目标无关的信息。
2. 提升模型的鲁棒性，降低模型对图像的敏感度。当训练数据都处于比较理想的状态时，如果碰到一些特殊情况，类似于遮挡、亮度不均、模糊等特殊情况很容易出现识别错误，对训练数据进行数据增强，如加上噪声、掩码等方法，可以有效提升模型的鲁棒性。
3. 增加训练数据量。提高模型泛化能力。
4. 避免样本不均衡。在工业缺陷检测、医疗疾病识别等方面，很容易出现正负样本极度不平衡的情况，通过对少样本进行一些数据增强，可以有效缓解样本不均衡比例。

继续看看定义在augmentions.py文件中的COCODetection的构造函数：

1. **def** \_\_init\_\_(self, image\_path, info\_file, transform=None,
2. target\_transform=None,
3. dataset\_name='MS COCO', has\_gt=True):
4. # Do this here because we have too many things named COCO
5. **from** pycocotools.coco **import** COCO
7. **if** target\_transform **is** None:
8. target\_transform = COCOAnnotationTransform()
10. self.root = image\_path
11. self.coco = COCO(info\_file) #将标签文件导入COCO API
13. #self.coco.imgToAnns 里面包含了标签文件中所有的bbox、category\_id、image\_id、segmentation信息
14. self.ids = list(self.coco.imgToAnns.keys())
15. **if** len(self.ids) == 0 **or** **not** has\_gt:
16. self.ids = list(self.coco.imgs.keys())
18. self.transform = transform
19. #transform是SSDAugmentation的实例对象
20. #COCOAnnotationTransform这个类作用：将COCO的标签转换成bbox coords and label index的张量
21. self.target\_transform = COCOAnnotationTransform()
23. self.name = dataset\_name
24. self.has\_gt = has\_gt

关注这里的COCOAnnotationTransform()，这个类的\_\_call\_\_方法将coco标签中的bbox和category\_id信息存到一个列表中：[xmin, ymin, xmax, ymax, category\_id]。这里需要注意coco标签中的bbox信息：[xmin, ymin, w, h]。

回到train.py文件中创建yolact算法模型对象的地方。

1. yolact\_net = Yolact()
2. net = yolact\_net
3. net.train()

Yolact()类定义在yolact.py文件中，其定义中默认以ResNet101作为backbone。

**3.3.3 训练过程参数的影响**

在深度学习神经网络中，超参数的调整至关重要，通过观察在训练过程中的监测指标如损失loss和准确率来判断当前模型处于什么样的训练状态，及时调整超参数从而更科学地训练模型能够提高训练效率和资源利用率。以下将分别对YOLACT算法使用的不同超参数及其影响进行介绍。

1. 学习率（learning rate或作lr）。学习率是指在优化算法中更新网络权重的幅度大小。学习率可以是恒定的、逐渐降低的，基于动量的或者是自适应的。不同的优化算法决定不同的学习率，学习率过大时则可能导致模型不收敛，损失loss不断上下震荡；学习率过小则可能导致模型收敛速度偏慢，需要花更长的时间进行训练。
2. 批次大小（batch\_size）。批次大小是指每一次训练神经网络送入模型的样本数，在卷积神经网络中，大批次通常可以使网络更快收敛，但由于内存资源的限制，批次过大可能会导致内存不够用或者程序内核崩溃。
3. 迭代次数。迭代次数是指整个训练集输入到神经网络进行训练的次数，当测试错误率和训练错误率相差较小时，可认为当前迭代次数合适；当测试错误率先变小后变大时则说明迭代次数过大了，需要减小迭代次数，否则容易出现过拟合。
4. 激活函数。在神经网络中，激活函数的作用是给神经网络加入一些非线性因素，从而使得神经网络可以更好地解决较为复杂的问题。
5. **图像语义分割的实际应用与细节优化**

**4.1 针对嵌套圆管的数据集制作**

数据图像为包括1到N个目标对象的圆管堆图像，由80%训练集和20%验证集组成，测试图像为训练图像经过图像数据增强得到。本实验使用labelme图像标注软件对原始图像进行实例分割和分类标注处理，转化为COCO数据集格式后，作为第一阶段的数据输入，处理后图像示例如图4-1所示：

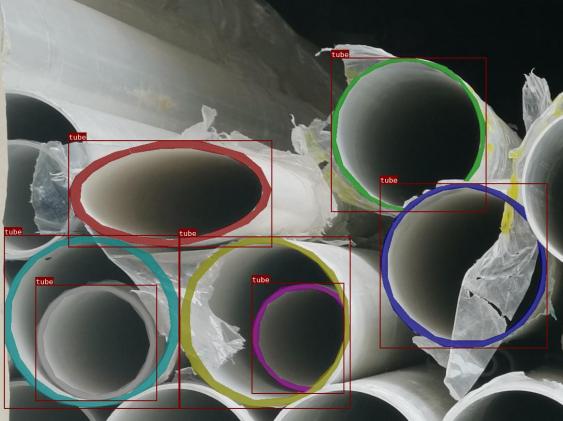


图4-1 原始图像（左），实例标签可视化图像（右）

**4.2 实验环境和参数设置**

本文实验环境基于Ubuntu 18.04.3 LTS，内核为GNU/Linux 4.15.0-171-generic x86\_64，使用Anaconda封装环境，python版本为3.7.11，深度学习库pytorch版本为1.11.0，torchvision版本为0.12.0，GPU驱动cudatoolkit版本为11.3.1。任何版本上的更改都会导致难以预料的兼容性问题，因此实验复现需要严格参照上述第三方库版本。

参数设置方面允许对数据集--dataset，批次大小--batch\_size，学习率--learning\_rate，循环周期--epoch等参数进行调整。此外，在读取图片的同时对其进行了次序相同、参数相等的增强处理，变换包括平移、旋转、缩放等基本变化，以及锐化、灰度增强、直方图匹配等增强方法。

**4.3 应用过程中的细节优化**

**4.4 效果展示与结果分析**

用于评估语义分割算法性能的标准指标是平均 IOU（Intersection Over Union，即交并比），IOU 定义如下：

这样的评价指标可以判断目标的捕获程度（使预测标签与标注尽可能重合），也可以判断模型的精确程度（使并集尽可能重合）。

**第五章 总结与展望**

**5.1 工作总结**

**5.2 未来展望**

参考文献：

1. <https://arxiv.org/pdf/1311.2524.pdf>
2. <https://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Girshick_Fast_R-CNN_ICCV_2015_paper.pdf>
3. <https://arxiv.org/pdf/1506.01497.pdf>

<https://openaccess.thecvf.com/content_ICCV_2017/papers/He_Mask_R-CNN_ICCV_2017_paper.pdf>