实例分割研究综述总结

CV开发者都爱看的 极市平台 2022-12-03 22:00:17 发表于广东 手机阅读

作者丨youtober@知乎（已授权）

来源丨https://zhuanlan.zhihu.com/p/412675982

编辑丨极市平台

极市导读

本文综述基于实例分割的最新进展和发展历程，首先介绍了实例分割的基本逻辑,总结了目前主要研究方法及其原理和网络架构，对已发表的主流实例分割方法进行分析，最后对实例分割任务目前面临 的问题以及未来的发展趋势做出了分析,并针对所面临的问题提出了一些切实可行的解决思路。 >>加入极市CV技术交流群，走在计算机视觉的最前沿

摘要

在计算机视觉领域，实例分割是一个很重要的研究主题，在地理信息系统、医学影像、自动驾驶、机器人等领域有着很重要的应用技术支持作用，具有十分重要的研究意义。本文综述基于实例分割的最新进展和发展历程，首先介绍了实例分割的基本逻辑,总结了目前主要研究方法及其原理和网络架构，对已发表的主流实例分割方法进行分析，最后对实例分割任务目前面临的问题以及未来的发展趋势做出了分析,并针对所面临的问题提出了一些切实可行的解决思路。

关键词 实例分割 图像分割 语义分割 深度学习

1. 介绍

图像分割是是指根据灰度、彩色、空间纹理、几何形状等特征把图像划分成若干个互不相交的区域，使得这些特征在同一区域内表现出一致性或相似性，而在不同区域间表现出明显的不同。如下图所示。

目标检测是识别图像中存在的内容和检测其位置，如下图，以识别和检测人（person）为例。

语义分割是对图像中的每个像素打上类别标签进行分类。如下图所示。

实例分割是目标检测和语义分割的结合，在图像中将目标检测出来（目标检测），然后对每个像素打上标签（语义分割）。如下图所示。

实例分割目的是将输入图像中的目标检测出来,并且对目标的每个像素分配类别标签.实例分割能够对前景语义类别相同的不同实例进行区分,这是它与语义分割的最大区别.相比语义分割,实例分割发展较晚,因此实例分割模型主要基于深度学习技术, 但它也是图像分割一个重要的组成部分.随着深度学习的发展,实例分割相继出现了 SDS(Simultaneous detection and segmentation)、DeepMask、MultiPath network 等方法,分割精度和效率逐渐得到提升。

1.1实例分割目前存在的一些问题和难点。

a.小物体分割问题。深层的神经网络一般有更大的感受野，对姿态，形变，光照等更具有鲁棒性，但是分辨率（ resolution）比较低，细节也丢失了；浅层的神经网络的感受野比较窄，细节比较丰富，分辨率比较大，但缺少了语义上的信息。因此，如果一个物体比较小时，它的细

节在浅层的CNN层中会更少，同样的细节在深层网络中几乎会消失。解决这个问题的方法有dilated convolution和增大特征的分辨率。

b.处理几何变换（ geometric transformation）的问题。对于几何变换，CNN本质上并不是空间不变的（ spatially invariant）。

c.处理遮挡（ occlusions）问题。遮挡会造成目标信息的丢失。目前提出了一些方法来解决这个问题，如 deformable ROI pooling，deformable convolution和 adversarial network。另外，也可能可以使用GAN来解决这个问题。

d.处理图像退化（ image degradations）的问题。造成图像退化的原因有光照，低质量的摄像机和图像压缩等。不过目前大多数数据集（如ImageNet，COCO和PASCAL VOC等）都不存在图像退化的问题。

2.实例分割的基本流程

实例分割模型一般由三部分组成: 图像输入、 实例分割处理、分割结果输出. 图像输入后, 模型一 般使用 VGGNet、ResNet 等骨干网络提取图像特 征, 然后通过实例分割模型进行处理. 模型中可以 先通过目标检测判定目标实例的位置和类别, 然后 在所选定区域位置进行分割, 或者先执行语义分割 任务, 再区分不同的实例, 最后输出实例分割结果。

2.1 实例分割的主要技术路线

实例分割的研究长期以来都有着两条线，分别是自下而上的基于语义分割的方法和自上而下的基于检测的方法，这两种方法都属于两阶段的方法。

自上而下的实例分割方法

思路是：首先通过目标检测的方法找出实例所在的区域（bounding box），再在检测框内进行语义分割，每个分割结果都作为一个不同的实例输出。通常先检测后分割，如FCIS, Mask-RCNN, PANet, Mask Scoring R-CNN；

自上而下的密集实例分割的开山鼻祖是DeepMask，它通过滑动窗口的方法，在每个空间区域上都预测一个mask proposal。这个方法存在以下三个缺点：

- mask与特征的联系（局部一致性）丢失了，如DeepMask中使用全连接网络去提取mask
- 特征的提取表示是冗余的， 如DeepMask对每个前景特征都会去提取一次mask
- 下采样（使用步长大于1的卷积）导致的位置信息丢失

自下而上的实例分割方法

将每个实例看成一个类别；然后按照聚类的思路，最大类间距，最小类内距，对每个像素做embedding，最后做grouping分出不同的instance。Grouping的方法：一般bottom-up效果差于top-down；

思路是：首先进行像素级别的语义分割，再通过聚类、度量学习等手段区分不同的实例。这种方法虽然保持了更好的低层特征（细节信息和位置信息），但也存在以下缺点：

- 对密集分割的质量要求很高，会导致非最优的分割
- 泛化能力较差，无法应对类别多的复杂场景
- 后处理方法繁琐

单阶段实例分割（Single Shot Instance Segmentation），这方面工作其实也是受到了单阶段目标检测研究的影响，因此也有两种思路，一种是受one-stage, anchor-based 检测模型如YOLO，RetinaNet启发，代表作有YOLACT和SOLO；一种是受anchor-free检测模型如 FCOS 启发，代表作有PolarMask和AdaptIS。

2.2 实例分割方法的发展历程

主要脉络图如下：

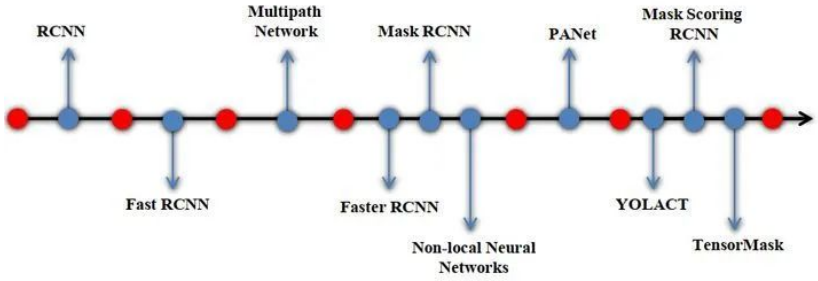


Figure 6. Timeline for notable techniques in instance segmentation

2.3 实例分割方法主要网络架构方法分类

主要有四种：掩模建议分类法、先检测再分割法、标记像素后聚类法和密集滑动窗口法

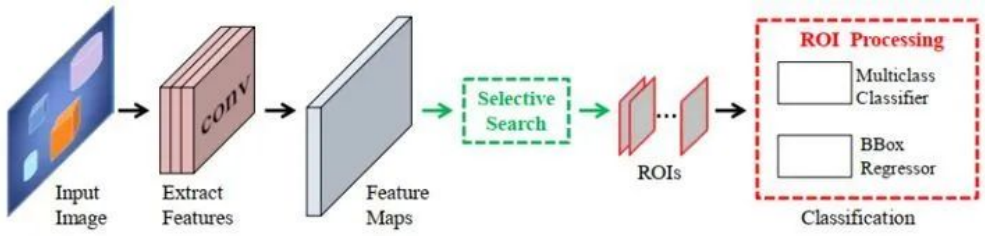
其对应英文名称和包含的主要技术方法如下表所示

Group	Technique
Classification of mask proposals	RCNN, Fast RCNN, Faster RCNN
Detection followed by segmentation	HTC, PANet, Mask RCNN, Mask Scoring RCNN, MPN, YOLACT
Labelling pixels followed by clustering	Deep Watershed Transform, Instance Cut
Dense sliding window methods	Deep Mask, Instance FCN, Tensor Mask

这四类方法的优缺点如下：

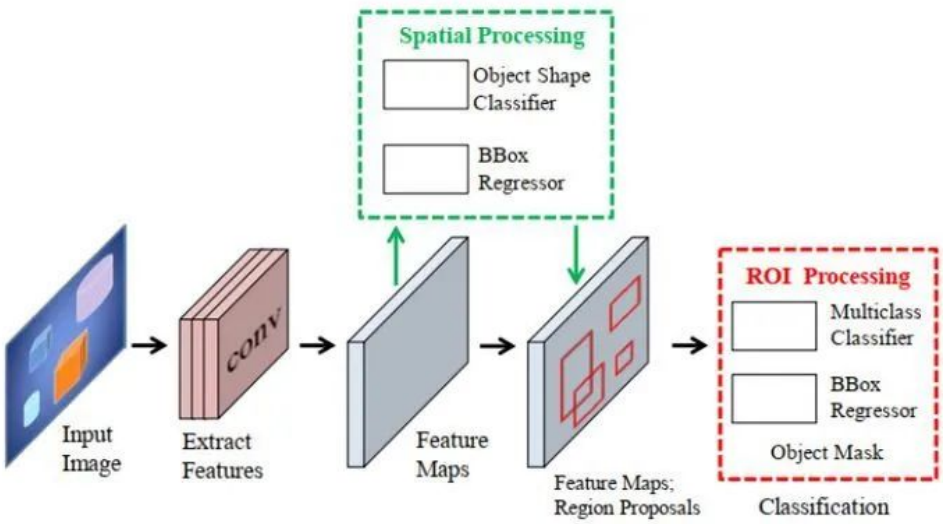
Group	Strengths	Weaknesses
Classification of mask proposals	✓ Relatively simple to implement; ✓ Modest segmentation accuracy;	○ Slow and difficult to optimize training; ○ Storage, time and detection-scale issues during training; ○ Slow testing; ○ Not suited for real time applications;
Detection followed by segmentation	✓ Relatively simple to train; ✓ better generalization; ✓ relatively faster (e.g. YOLACT); ✓ good segmentation accuracy;	○ Depend on a complicated training pipeline which is difficult to train, and to optimize;
Labelling pixels followed by clustering	✓ Use some recently investigated techniques; ✓ Relatively simpler techniques;	○ Lesser segmentation accuracy; ○ Intense computation necessitates high computational power; ○ Not suited for real time applications;
Dense sliding window methods	✓ Relatively unexplored area; ✓ Modest segmentation accuracy;	○ Use complex algorithms; ○ Difficult to train and optimize; ○ Not suited for real time applications;

2.3.1 掩模建议分类法



General framework for Classification for Mask Proposals Techniques

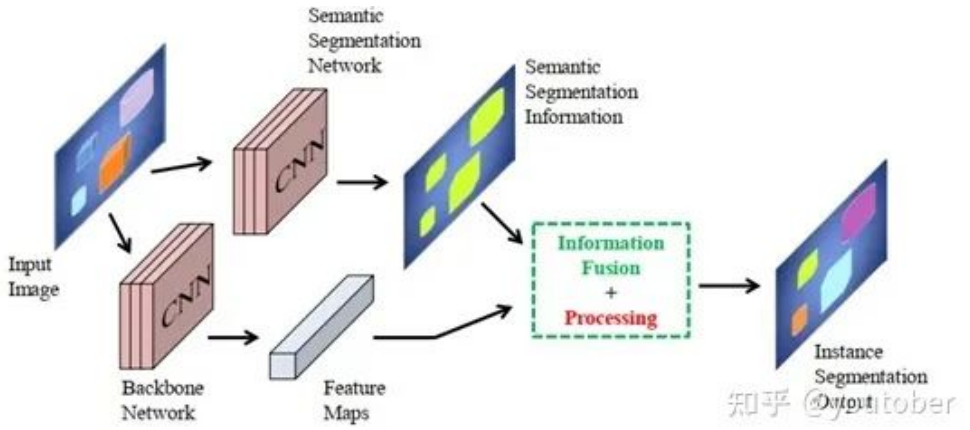
2.3.2 先检测再分割法



General framework for Detection Followed by Segmentation

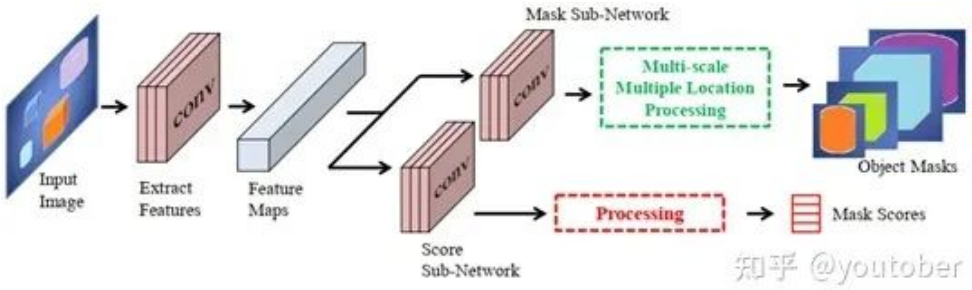
2.3.3 标记像素后聚类法

该方法受益于语义分割，可以预测高分辨率的对象掩模。与分割检测跟踪技术相比，标签像素跟踪聚类方法在经常使用的基准上精度较低。由于像素标记需要密集的计算，通常需要更多的计算能力。



General framework for Labelling Pixels Followed by Clustering Techniques

2.3.4 密集滑动窗口法



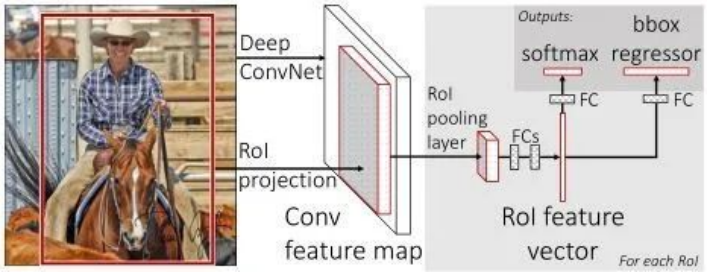
General framework for Dense Sliding Window Methods

3 实例分割的典型方法

3.1 DeepMask

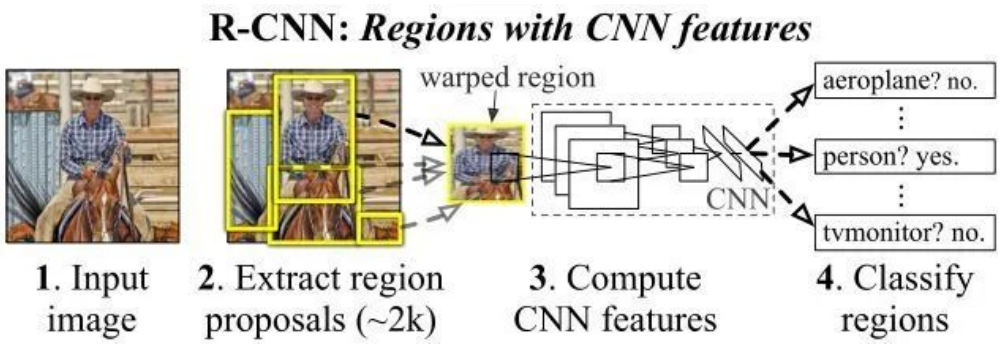
DeepMask 网络采用 VGGNet 对输入图像提 取特征, 生成分割提议, 提取的特征为两个分支所共享, 第 1 个分支对选中的物体预测一个分割掩码, 第 2 个分支对输入的 Patch 预测一个目标得分. 该 网络在 PASCAL VOC 2007 和 MS COCO 数据集 上进行了验证, 分割精度良好。

3．2 Fast-CNN



Fast RCNN解决了RCNN的一些问题，从而提高了目标检测能力。Fast RCNN使用检测器的端到端训练。它通过同时学习softmax分类器和类特定的BBox回归简化了训练过程，而不是像RCNN那样单独训练模型的各个组件。快速RCNN共享区域方案的卷积计算，然后在最后一个卷积层和第一个全连接层之间添加一个ROI池化层，提取每个区域方案的特征。聚类利用特征层扭曲的概念来实现图像层扭曲。将ROI池化层特征分解为一组全连通层，最后分解为目标类别预测软最大概率和类别建议精细化偏移量两层。与RCNN相比，Fast RCNN在很大程度上提高了效率，训练速度提高了3倍，测试速度提高了10倍。

3.3 Mask R-CNN (2017.3)



Mask R-CNN 由 He 等[39] 提出, 是在 Faster RCNN[40] 基础上扩展而来的一种全新的实例分割模 型. Mask R-CNN 属于两阶段方法, 第 1 阶段使用RPN (Region proposal network) 来产生 ROI (Region of interest) 候选区域. 第 2 阶段模型对每个 ROI 的类别、边界框偏移和二值化掩码进行预测. 掩码由新增加的第 3 个分支进行预测, 这是 Mask R-CNN 与其他方法的不

同点. 此外, Mask R-CNN 提出了 ROIAlign, 在下采样时对像素进行对准, 使 得分割的实例位置更加准确.

RCNN集成了AlexNet和使用选择性搜索技术的区域方案。RCNN模型的训练包括以下步骤。第一步涉及计算使用选择性搜索获得的类不可知区域建议。下一步是CNN模型微调，包括使用区域建议微调预先训练的CNN模型，如AlexNet。接下来，利用CNN提取的特征来训练一组类特异性支持向量机(SVM)分类器，该分类器取代了通过微调学习的softmax分类器。然后使用CNN获得的特征对每个对象类进行类特异性边界盒回归训练。

其结构与Faster RCNN非常类似，但有3点主要区别：

- 在基础网络中采用了较为优秀的ResNet-FPN结构，多层特征图有利于多尺度物体及小物体的检测。原始的FPN会输出P2、P3、P4与P54个阶段的特征图，但在Mask RCNN中又增加了一个P6。将P5进行最大值池化即可得到P6，目的是获得更大感受野的特征，该阶段仅仅用在RPN网络中。
- 提出了RoI Align方法来替代RoI Pooling，原因是RoI Pooling的取整做法损失了一些精度，而这对于分割任务来说较为致命。 Maks RCNN提出的RoI Align取消了取整操作，而是保留所有的浮点，然后通过双线性插值的方法获得多个采样点的值，再将多个采样点进行最大值的池化，即可得到该点最终的值。
- 得到感兴趣区域的特征后，在原来分类与回归的基础上，增加了一个Mask分支来预测每一个像素的类别。具体实现时，采用了FCN（Fully Convolutional Network）的网络结构，利用卷积与反卷积构建端到端的网络，最后对每一个像素分类，实现了较好的分割效果。

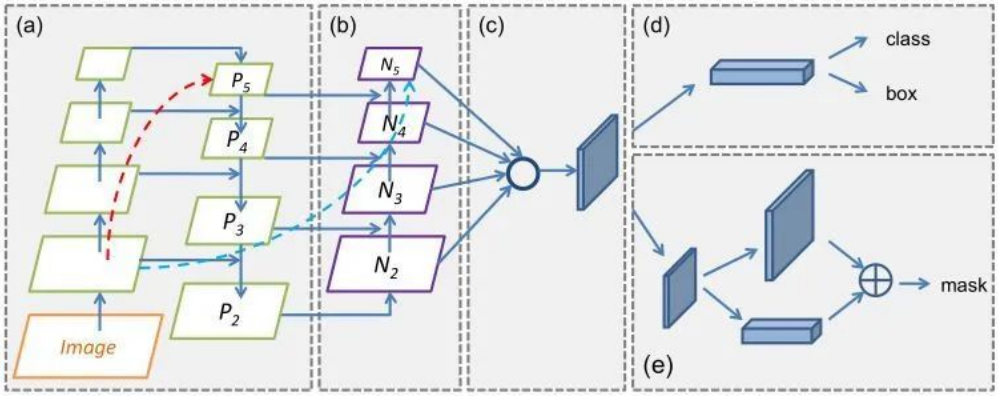
Mask R-CNN算法的主要步骤为：

- 首先，将输入图片送入到特征提取网络得到特征图。
- 然后对特征图的每一个像素位置设定固定个数的ROI（也可以叫Anchor），然后将ROI区域送入RPN网络进行二分类(前景和背景)以及坐标回归，以获得精炼后的ROI区域。
- 对上个步骤中获得的ROI区域执行论文提出的ROIAlign操作，即先将原图和feature map的pixel对应起来，然后将feature map和固定的feature对应起来。
- 最后对这些ROI区域进行多类别分类，候选框回归和引入FCN生成Mask，完成分割任务。

总的来说，在Faster R-CNN和FPN的加持下，Mask R-CNN开启了R-CNN结构下多任务学习的序幕。它出现的时间比其他的一些实例分割方法（例如FCIS）要晚，但是依然让proposal-based instance segmentation的方式占据了主导地位（尽管先检测后分割的逻辑不是那么地自然）。

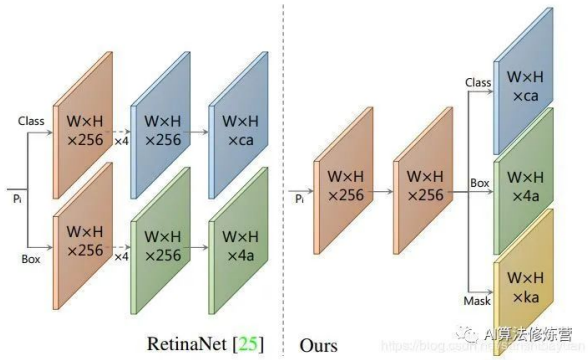
Mask R-CNN利用R-CNN得到的物体框来区分各个实例，然后针对各个物体框对其中的实例进行分割。显而易见的问题便是，如果框不准，分割结果也会不准。因此对于一些边缘精度要求高的任务而言，这不是一个较好的方案。同时由于依赖框的准确性，这也容易导致一些非方正的物体效果比较差。

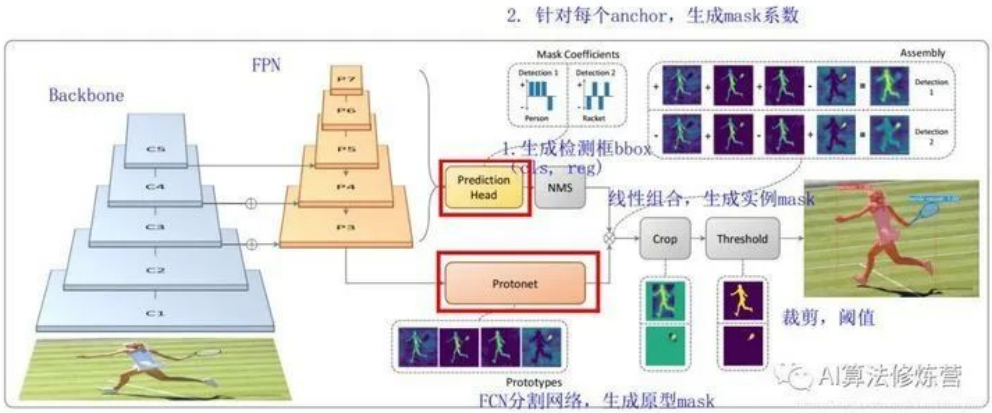
3.4 PANet



PANet 是 Liu 等[41] 提出的一种两阶段实例分 割模型. 为了缩短信息通路, 该模型利用低层精确 的定位信息提升特征金字塔, 创建了自底向上的路 径增强. 为了恢复候选区域和所有特征层之间被破 坏的信息通路, Liu 等[41] 开发了自适应特征池化, 用 来汇聚每个候选区域所有特征层的特征. 此外, 模 型用全连接层来增强掩码预测, 由于具有全卷积网 络的互补特性, 模型获得了每个候选区域的不同视 图. 由于这些改进, PANet 在 MS COCO 2017 实 例分割任务上排名第一. 提出了一种用于实例分割任务的基于框架，旨在提高信息的流动。改进了深层网络的特征层次，在底层使用与定位相关的特定信号。这个过程称为自底向上路径增强。它使得底层和深层网络顶层特征之间的信息路径更短。还提出了一种被称为自适应特性池的技术，它将特征网格和所有层次的特征联系起来。由于这种技术，在每一级特征的相关信息流到后续子网络用于产生建议。一个备用的分支段捕获各种提议视图，以增强生成掩码的预测

3.5 YOLCAT（2019.4）





- YOLACT将掩模分支添加到现有的一阶段（one-stage）目标检测模型，其方式与Mask R-CNN对 Faster-CNN 操作相同，但没有明确的定位步骤。
- YOLACT将实例分割任务拆分成两个并行的子任务：（1）通过一个Protonet网络， 为每张图片生成 k 个 原型mask;（2）对每个实例，预测k个的线性组合系数（Mask Coefficients）。最后通过线性组合，生成实例mask，在此过程中，网络学会了如何定位不同位置、颜色和语义实例的mask。

- YOLACT将问题分解为两个并行的部分，利用 fc层（擅长产生语义向量）和 conv层（擅长产生空间相干掩模）来分别产生“掩模系数”和“原型掩模”。然后，因为原型和掩模系数可以独立地计算，所以 backbone 检测器的计算开销主要来自合成（assembly）步骤，其可以实现为单个矩阵乘法。通过这种方式，我们可以在特征空间中保持空间一致性，同时仍然是一阶段和快速的。

Backbone：Resnet 101+FPN，与RetinaNet相同；**Protonet**：接在FPN输出的后面，是一个FCN网络，预测得到针对原图的原型mask；**Prediction Head**：相比RetinaNet的Head，多了一个Mask Cofficient分支，预测Mask系数，因此输出是4*c+k。

可以看到head上增加了一支mask系数分支用于将prototypes进行组合得到mask的结果。当然按NMS的位置看，其同样需要有bbox的准确预测才行，并且该流程里不太适合用soft NMS进行替代。需要注意的是，在训练过程中，其用groundtruth bbox对组合后的全图分割结果进行截取，再与groundtruth mask计算损失。这同样需要bbox结果在前作为前提，以缓解前后景的像素不均衡情况。

至于后续的YOLCAT++，则主要是加入了mask rescoring的概念和DCN结构，进一步提升精度。（1）参考Mask Scoring RCNN，添加fast mask re-scoring分支，更好地评价实例mask的好坏；（2）Backbone网络中引入可变形卷积DCN；（3）优化了Prediction Head中的anchor设计。

3.6 PolarMask (2019.10)

PolarMask则是进一步细化了边界的描述，使得其能够适应mask的问题。 PolarMask最重要的特点是：(1) anchor free and bbox free，不需要出检测框；(2) fully convolutional network, 相比FCOS把4根射线散发到36根射线，将instance segmentation和object detection用同一种建模方式来表达。

PolarMask 基于极坐标系建模轮廓，把实例分割问题转化为实例中心点分类(instance center classification)问题和密集距离回归(dense distance regression)问题。 同时，我们还提出了两个有效的方法，用来优化high-quality正样本采样和dense distance regression的损失函数优化，分别是Polar CenterNess和 Polar IoU Loss。没有使用任何trick(多尺度训练，延长训练时间等)，PolarMask 在ResNext 101的配置下 在coco test-dev上取得了32.9的mAP。这是首次，证明了更复杂的实例分割问题，可以在网络设计和计算复杂度上，和anchor free物体检测一样简单。

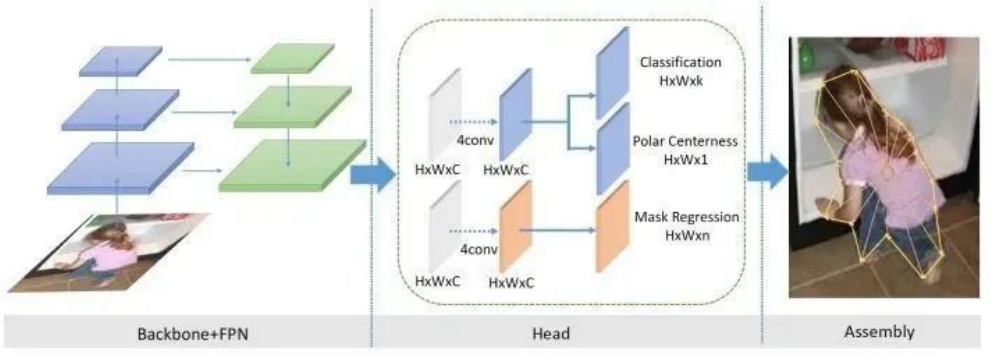


Figure 2 – The overall pipeline of PolarMask. The left part contains the backbone and feature pyramid to extract feature maps. The middle part is the two heads for classification and polar mask regression. H, W, C are the height, width, channels of feature maps, k is the number of categories (e.g., $k = 80$ in COCO dataset), n is the number of rays (e.g., $n = 36$)

PolarMask网络架构图

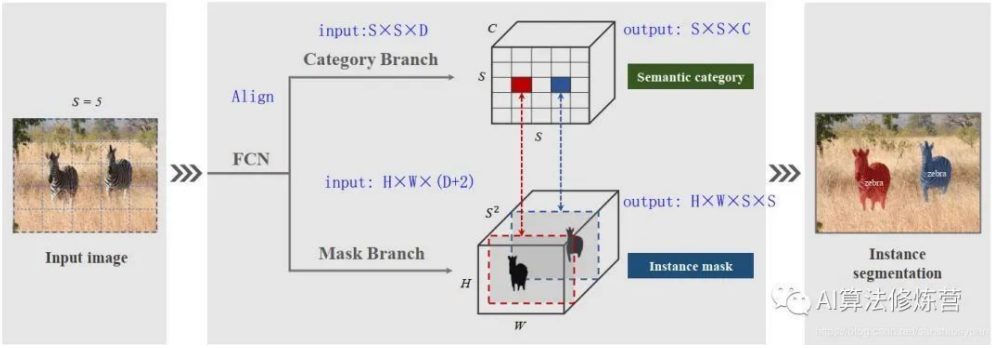
整个网络和FCOS一样简单，首先是标准的backbone + fpn模型，其次是head部分，我们把fcos的bbox分支替换为mask分支，仅仅是把channel=4替换为channel=n, 这里n=36，相当于36根射线的长度。同时我们提出了一种新的Polar Centerness 用来替换FCOS的bbox centerness。可以看到，在网络复杂度上，PolarMask和FCOS并无明显差别。

3.7 SOLO (2019.12)

SOLO将一张图片划分S×S的网格，这就有了S*S个位置。不同于TensorMask和DeepMask将mask放在了特征图的channel维度上，SOLO参照语义分割，将定义的物体中心位置的类别放在了channel维度上，这样就保留了几何结构上的信息。

本质上来说，一个实例类别可以去近似一个实例的中心的位置。因此，通过将每个像素分类到对应的实例类别，就相当于逐像素地回归出物体的中心、这就将一个位置预测的问题从回归的问题转化成了分类的问题。这么做的意义是，分类问题能够更加直观和简单地用固定的channel数、同时不依赖后处理方法（如分组和学习像素嵌入embedding）对数量不定的实例进行建模。

对于尺寸的处理，SOLO使用了FPN来将不同尺寸的物体分配到不同层级的特征图上，依次作为物体的尺寸类别。这样，所有的实例都被分别开来，就可以去使用实例类别去分类物体了。



SOLO网络架构图

SOLO将图片划分成S×S的网格，如果物体的中心（质心）落在了某个网格中，那么该网格就有了两个任务：（1）负责预测该物体语义类别（2）负责预测该物体的instance mask。这就对应了网络的两个分支Category Branch和Mask Branch。同时，SOLO在骨干网络后面使用了FPN，用来应对尺寸。FPN的每一层后都接上述两个并行的分支，进行类别和位置的预测，每个分支的网格数目也相应不同，小的实例对应更多的的网格。

Category Branch：Category Branch负责预测物体的语义类别，每个网格预测类别S×S×C，这部分跟YOLO是类似的。输入为Align后的S×S×C的网格图像，输出为S×S×C的类别。这个分支使用的损失函数是**focal loss**。

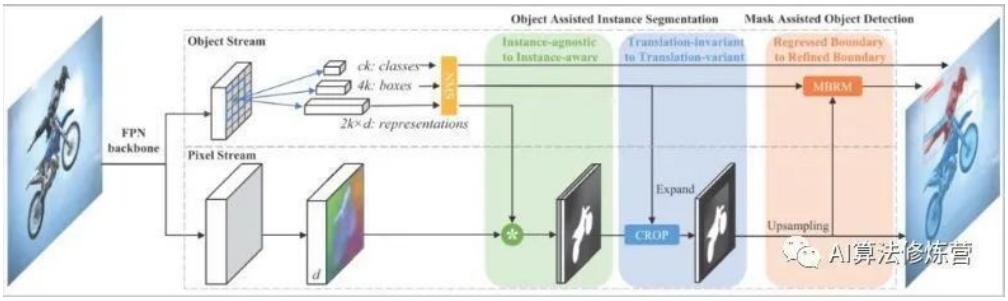
Mask Branch：预测instance mask的一个直观方法是类似语义分割使用FCN，但FCN是具有空间不变性（spatiallly invariant）的，而我们这边需要位置上的信息。因此，作者使用了CoordConv，将像素横纵坐标x，y（归一化到[-1,1]）与输入特征做了concat再输入网络中。这样输入的维度就是 H*W*(D+2）了。

实验结果：

	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
two-stage:							
MNC [3]	Res-101-C4	24.6	44.3	24.8	4.7	25.9	43.6
FCIS [10]	Res-101-C5	29.2	49.5	–	7.1	31.3	50.0
Mask R-CNN [7]	Res-101-FPN	35.7	58.0	37.8	15.5	38.1	52.4
Mask R-CNN* [2]	Res-50-FPN	36.8	59.2	39.3	17.1	38.7	52.1
Mask R-CNN* [2]	Res-101-FPN	38.3	61.2	40.8	18.2	40.6	54.1
one-stage:							
TensorMask [2]	Res-50-FPN	35.4	57.2	37.3	16.3	36.8	49.3
TensorMask [2]	Res-101-FPN	37.1	59.3	39.4	17.4	39.1	51.6
YOLACT [1]	Res-101-FPN	31.2	50.6	32.8	12.1	33.3	47.1
PolarMask [27]	Res-101-FPN	30.4	51.9	31.0	13.4	32.4	42.8
ours:							
SOLO	Res-50-FPN	36.8	58.6	39.0	15.9	39.5	52.1
SOLO	Res-101-FPN	37.8	59.5	40.4	16.4	41.4	54.5
SOLO	Res-DCN-101-FPN	40.4	62.7	43.3	17.6	43.3	58.9

SOLO的精度已经超越了Mask R-CNN，相较思路类似的PolarMask也有较大的优势。

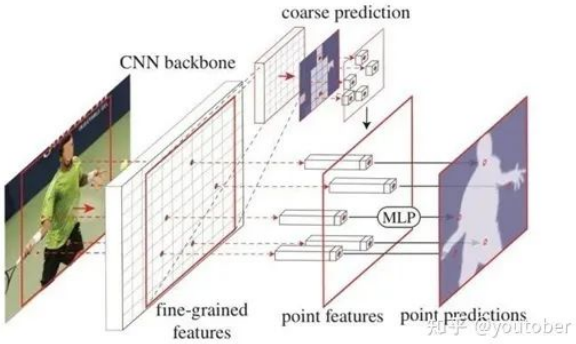
3.8 RDSNet（2019.12）



RDSNet方法的出发点是检测阻碍不应该成为分割效果的阻碍，两种应该循环相互促进。有可能存在的情况是分割本身是比较准确的，但是因为定位不准，导致分割结果也比较差；这时候如果能提前知道分割的结果，那么检测的结果也会更好些。

有用到YOLCAT的方式，去获得提取获取分割结果。当然这里从embedding的角度出发，还结合了前后景的处理（实验中说明前后景correlation比单前景linear combination要好）。得到box预测结果后是需要进行NMS，以及expand操作的，以确保尽可能多的有效区域被选进来（训练时1.5，测试时1.2）。之后再通过Mask-based Boundary Refinement模块对物体的边框进行调整。

3.9 PointRend (2019.12)



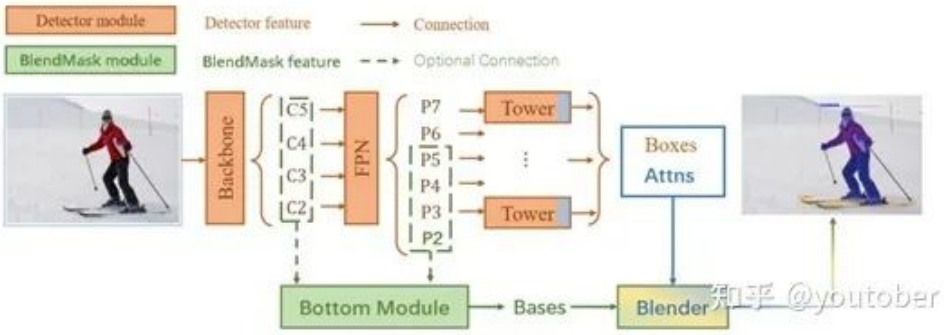
PointRend借鉴了Render的思想，在尺度方式变化时由于采样的方式（不是连续坐标的设定吗），使得锯齿现象不会很明显。因此PointRend是利用一种非均匀采样的方式来确定在分辨率提高的情况下，如何确定边界上的点，并对这些点归属进行判别。 本质上其实是一个新型上采样方法，针对物体边缘的图像分割进行优化，使其在难以分割的物体边缘部分有更好的表现。

PointRend 方法要点总结来说是一个迭代上采样的过程：

while 输出的分辨率 < 图片分辨率：

1. 对输出结果进行2倍双线性插值上采样得到 coarse prediction_i。
2. 挑选出 N 个“难点”，即结果很有可能和周围点不一样的点（例如物体边缘）。
3. 对于每个难点，获取其“表征向量”，“表征向量”由两个部分组成，其一是低层特征（fine-grained features），通过使用点的坐标，在低层的特征图上进行双线性插值获得（类似 RoI Align），其二是高层特征（coarse prediction），由步骤 1 获得。
4. 使用 MLP 对“表征向量”计算得到新的预测，更新 coarse prediction_i 得到 coarse prediction_i+1。这个 MLP 其实可以看做一个只对“难点”的“表征向量”进行运算的由多个 conv1x1 组成的小网络。

3.10 BlendMask (2021.1)



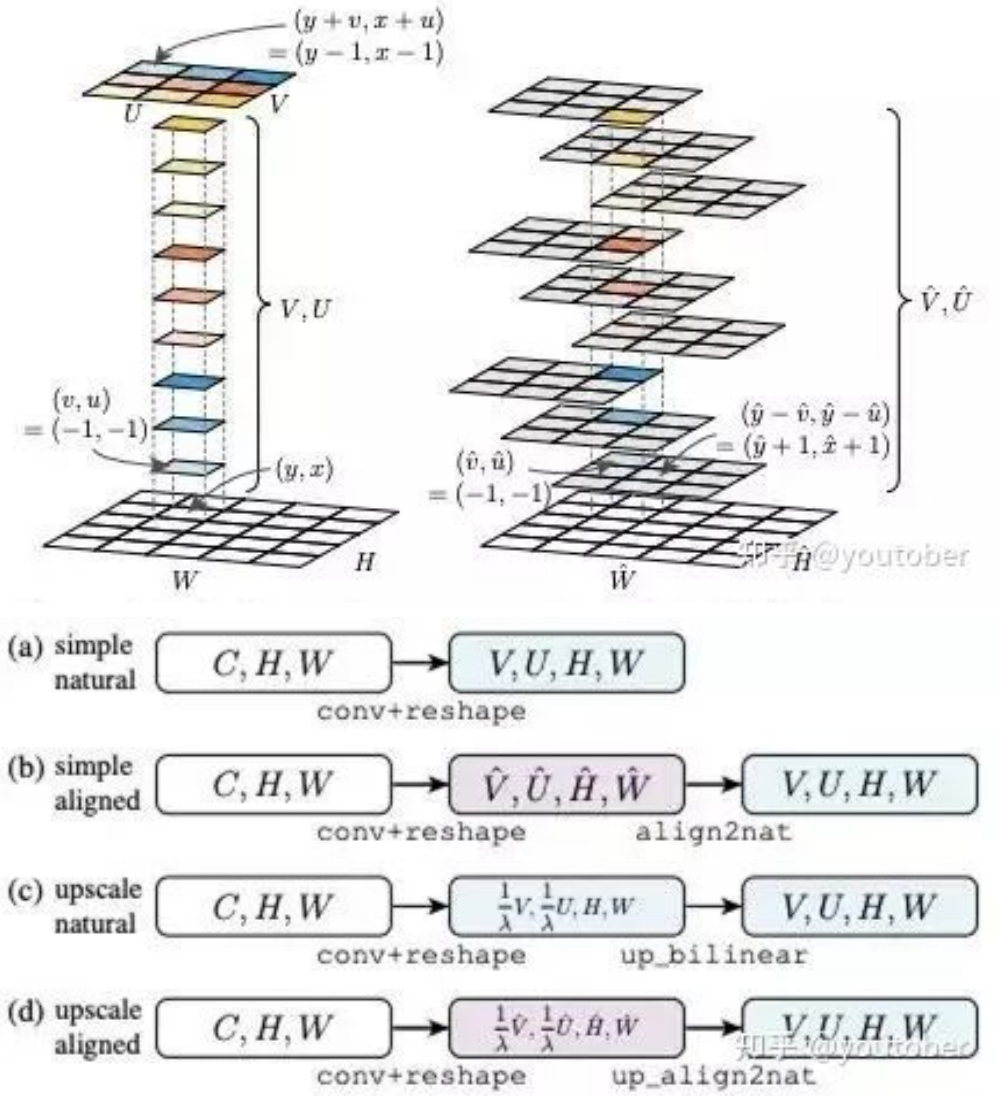
BlendMask是一阶段的密集实例分割方法，结合了Top-down和Bottom-up的方法的思路。它通过在anchor-free检测模型FCOS的基础上增加了Bottom Module提取low-level的细节特征，并在instance-level上预测一个attention；借鉴FCIS和YOLACT的融合方法，提出了Blender模块来更好地融合这两种特征。最终，BlendMask在COCO上的精度（41.3AP）与速度（BlendMask-RT 34.2mAP, 25FPS on 1080ti）都超越了Mask R-CNN。

detector module直接用的FCOS，BlendMask模块则由三部分组成：bottom module用来对底层特征进行处理，生成的score map称为Base；top layer串接在检测器的box head上，生成Base对应的top level attention；最后是blender来对Base和attention进行融合。

BlendMask 的**优势**：

- 计算量小：使用一阶段检测器FCOS，相比Mask R-CNN使用的RPN，省下了对position-sensitive feature map及mask feature的计算，
- 还是计算量小：提出attention guided blender模块来计算全局特征（global map representation），相比FCN和FCIS中使用的较复杂的hard alignment在相同分辨率的条件下，减少了十倍的计算量；
- mask质量更高：BlendMask属于密集像素预测的方法，输出的分辨率不会受到 top-level 采样的限制。在Mask R-CNN中，如果要得到更准确的mask特征，就必须增加RoI Pooler的分辨率，这样变回成倍增加head的计算时间和head的网络深度；
- 推理时间稳定：Mask R-CNN的推理时间随着检测的bbox数量增多而增多，BlendMask的推理速度更快且增加的时间可以忽略不计
- Flexible：可以加到其他检测算法里面

3.11 TensorMask



TensorMask将实例分割视为 4D 张量预测，TensorMask 表示的核心想法是使用结构化的 4D tensors 表示空间域上的 mask。

TensorMask 是一个 dense sliding-window 实例分割框架，首次在定性和定量上都接近于 Mask R-CNN 框架。TensorMask 为实例分割研究建立了一个概念互补的方向。

3.12 主要方法在在COCO数据集上的指标对比：

Method	AP	AP50	AP75	APs	APm	APL
FCIS	29.2	49.5		7.1	31.3	50.0
Mask R-CNN	37.1	60.0	39.4	16.9	39.9	53.5
YOLACT-700	31.2	50.6	32.8	12.1	33.3	47.1
PolarMask	32.9	55.4	33.8	15.5	35.1	46.3
SOLO	40.4	62.7	43.3	17.6	43.3	58.9
PointRend	40.9					
BlendMask	41.3	63.1	44.6	22.7	44.1	54.5

4.实例分割常用数据集

实例分割割常用数据集有 PASCAL VOC、MS COCO、Cityscapes、ADE20k 等.本小节从图像数、类别数、样本数等方面介绍几种 常用数据集.

4.1 PASCAL VOC 数据集

VOC数据集是计算机视觉主流数据集之一，可以作分类，分割，目标检测，动作检测和人物定位五类任务数据集，PASCAL VOC 在 2005 ~ 2012 年每年发布关 于图像分类、目标检测、图像分割等任务的子数据 集,并举行世界级的计算机视觉大赛.PASCAL VOC 数据集最初有 4 类,最后稳定在21 类,对于分割任务,这些类别有汽车、房屋、动物、飞机、自行车、船、 公共汽车、小汽车、摩托车、火车等,测试图像从早 期的 1578幅最后稳定在11 540 幅.PASCAL VOC 数据集包括训练集和测试集,对于实际比赛有一个 独立的测试集.2012 年以后 PASCAL VOC 大赛停办,但是数据集开源,可以下载使用.

4.2 Microsoft Common Objects in Context (MS COCO)

MS COCO是另一个大规模物体检测，分割及文字定位数据集。该数据集包含众多类别，以及大量的标签。它总共有91个物体类别,32.8万幅图像,超过8万幅 图像用于训练,4万多幅图像用于验证,8万多幅图 像用于测试,拥有250万个标注实例.MS COCO 数据集的每一类物体的图像数量多,标注精细,数 据场景多样性高,是目前比较流行的数据集.

4.3 Cityscapes

Cityscapes是另一个大规模数据集，其关注于城市街景的语义理解。它包含了一组来自50个城市的街景的不同的立体视频序列，有5k帧的高质量像素级标注，还有一组20k的弱标注帧。Cityscapes 数据集是一个城市街道场景的数据 集,拥有精细标注的5000 幅城市驾驶场景图像,其中 2975 幅用于训练,500幅用于验证,1525幅用于 测试,还有20000 幅粗标注的图像,一般使用精细 标注的那部分数据.该数据集包含来自50个城市 街道场景中记录的图像,是一个流行的街道场景数 据集.

4.4 ADE20K

ADE20K 是一个新的场景理解数据集,总共 有2万多幅图像,其中训练集有 20 210 幅图像,验证集有 2 000 幅图像,测试集有 3 352 幅图像,以开放字典标签集密集注释.ADE20K 包含 151 个物体 类别,如汽车、天空、街道、窗户、草坪、海面、咖啡桌 等,每幅图像可能包含多个不同类型的物体,物体尺度变化大,因此检测难度高

参考文献

- [1]实例分割最新最全面综述：从Mask R-CNN到BlendMask - 云+社区 - 腾讯云 (tencent.com)
- [2]一文读懂语义分割与实例分割 - 知乎 (zhihu.com)
- [3]实例分割最新最全面综述：从Mask R-CNN到BlendMask - 云+社区 - 腾讯云 (tencent.com)
- [4]https://zhuanlan.zhihu.com/p/165135767
- [5]http://www.aas.net.cn/cn/article/doi/10.16383/j.aas.c200657
- [6]何恺明等最新论文：实例分割全新方法 TensorMask，效果比肩 Mask R-CNN - 知乎 (zhihu.

公众号后台回复“CCF2022”2022（拟定）目录PDF下载～

极市平台
为计算机视觉开发者提供全流程算法开发训练平台，以及大咖技术分享、社区交流、竞...
848篇原创内容

公众号

极市干货

技术干货：数据可视化必须注意的30个小技巧总结丨如何高效实现矩阵乘？万文长字带你从CUDA初学者的角度入门

实操教程：Nvidia Jetson TX2使用TensorRT部署yolov5s模型丨基于YOLOV5的数据集标注&训练，Windows/Linux/Jetson Nano多平台部署全流程