目　　录

一、外文参考文献翻译-译文

摘　　要

我们提出了一个概念上简单，灵活和通用的对象实例分割框架。我们的方法同时有效地检测图像中的对象并为每个实例生成高质量的图像分割。这种方法成为 Mask R-CNN，它通过添加一个与现有边界框识别分支并行的 (Mask) 掩码检测分支扩展了 Faster R-CNN。Mask R-CNN 很容易训练，因为只是在 R-CNN 的基础上增加一小部分，可以以 5 fps 运行。此外 Mask R-CNN 易于推广到其他任务，例如，我们在同一框架中估计人体姿势，我们在 COCO 训练集的所有三个分支中都显示了最佳结果挑战，包括实例分割，边界框对象检测和人员关键点检测。不吹不夸，Mask R-CNN 优于现有每个任务的单一模型条目，包括 COCO 2016 挑战赛冠军。我们希望我们的简单和有效的方法将作为一个坚实的基线和帮助简化未来在实例级认可方面的研究。代码已在以下地址提供：https://github.com/facebookresearch/Detectron.

关键词：反射式，光纤，位移，测量

1 绪论

视觉社区已在短时间迅速改进了对象检测和短语期间的语义分割结果。在很大程度上，这些进步已经被强大的基线系统推动，例如 Fast / Faster RCNN 和完全卷积网络（FCN）分别对应对象检测和语义分割的框架。这些方法在概念上是直观的，并具有灵活性和稳健性，以及快速培训和推理时间。我们在这项工作中的目标是开发一个相对可行的实例分割框架。

实例细分具有挑战性，因为它需要同时正确检测图像中的所有对象并精确地分割每个实例。因此它结合了经典计算机视觉任务的目标检测和语义分段，前者的目标是使用边界框对单个对象进行分类和定位，后者的目标是将每个像素分类为一组固定的类别，不区分对象实例。鉴于此，人们可能会期待一种复杂的方法要取得好成绩。但是，我们提出了一个可以超越先前的最新实例分割结果的简单，灵活，快速令人惊讶的系统。

我们的方法称为 Mask R-CNN，它扩展了 Faster R-CNN 通过添加一个与现有用于分类和边界框回归的分支区域平行的分支来在每个感兴趣区域（RoI）预测分割掩码。（图 1）。掩码分支是应用于每个 RoI 的小 FCN，用于预测像素到像素中的分割掩模方式。在 Faster R-CNN 框架下 Mask R-CNN 易于实现和训练，这促进了广泛的灵活架构设计。另外，掩码分支只增加了一个小的计算开销，实现快速系统和快速实验。

原则上，Mask R-CNN 直观上是一个 Faster R-CNN 的扩展，但正确构建掩模分支对于取得好成绩至关重要。最重要的是，Faster R-CNN 不是为网络输入和输出间像素到像素之间的对齐而设计的。这一点在 RoIPool（解决事例核心操作的关键）如何执行用于特征提取的粗略空间量化中最为明显。为了解决这个错位，我们提出一个简单的，无量化的层，称为 RoIAlign，它忠实地保留确切的空间位置。尽管看似微不足道的变化，但 RoIAlign 产生了巨大的影响：掩模精度 提高了 10％到 50％，显示出更严格的本地化指标带来更大的收益。第二，我们发现解耦掩模和类别预测很重要：我们每个类独立地预测二进制掩码，而不是在各个类之间竞争，并依赖于网络的 RoI 分类分支来预测类别。相反，FCN 通常执行每像素多类别分类，它结合了分割和分类，并且基于我们的实验在例如细分方面效果不佳。

没有花里胡哨，Mask R-CNN 超越了所有 COCO 以前最先进的单一模型结果实例分割任务 [28]，包括重工程 2016 年比赛获胜者的参赛作品。作为副产品，我们的方法也擅

长 COCO 对象检测任务。在消融实验中，我们评估多个基本实例化，它允许我们展示它鲁棒性和分析核心因素的影响。我们的模型在 GPU 上每帧可以运行大约 200ms，COCO 的训练只需要 8 GPU 机器一到两天的时间。我们相信快速的训练和测试速度还有框架的灵活性和准确性会有利于和简化实例细分的未来研究。

最后，我们通过 COCO 数据集中人体关键点上的姿态估计任务展示了我们框架的一般性。通过将每个关键点视为一个热点二进制掩码，以最小修改 Mask R-CNN 即可应用于检测特定于实例的姿势。我们已经发布了代码以促进未来的研究。

二、外文参考文献翻译-原文

Mask R-CNN

Kaiming He Georgia Gkioxari Piotr Dollár Ross Girshick

Facebook AI Research (FAIR)

Abstract

We present a conceptually simple, flexible, and general framework for object instance segmentation. Our approach efficiently detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance. The method, called Mask R-CNN, extends Faster R-CNN by adding a branch for predicting an object mask in parallel *with the existing branch for bounding box recognition. Mask R-CNN is simple to train and adds only a small overhead to Faster R-CNN, running at 5 fps. Moreover, Mask R-CNN is easy to generalize to other tasks,* e.g.*, allowing us to estimate human poses in the same framework. We show top results in all three tracks of the COCO suite of challenges, including instance segmentation, bounding-box object detection, and person keypoint detection. Without bells and whistles, Mask R-CNN outperforms all existing, single-model entries on every task, including the COCO 2016 challenge winners. We hope our simple and effective approach will serve as a solid baseline and help ease future research in instance-level recognition. Code has been made available at:* https://github.com/facebookresearch/Detectron.

1. Introduction

The vision community has rapidly improved object detection and semantic segmentation results over a short period of time. In large part, these advances have been driven by powerful baseline systems, such as the Fast/Faster R-CNN [12, 36] and Fully Convolutional Network (FCN) [30] frameworks for object detection and semantic segmentation, respectively. These methods are conceptually intuitive and offer flexibility and robustness, together with fast training and inference time. Our goal in this work is to develop a comparably enabling framework for *instance segmentation*.

Instance segmentation is challenging because it requires the correct detection of all objects in an image while also precisely segmenting each instance. It therefore combines elements from the classical computer vision tasks of *object detection*, where the goal is to classify individual objects and localize each using a bounding box, and *semantic segmentation*, where the goal is to classify each pixel into a fixed set of categories without differentiating object instances.[1] Given this, one might expect a complex method is required to achieve good results. However, we show that a surprisingly simple, flexible, and fast system can surpass prior state-of-the-art instance segmentation results.

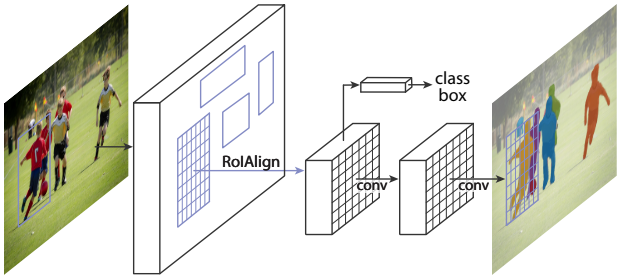

Figure 1. The **Mask R-CNN** framework for instance segmentation.

Our method, called *Mask R-CNN*, extends Faster R-CNN [36] by adding a branch for predicting segmentation masks on each Region of Interest (RoI), in *parallel* with the existing branch for classification and bounding box regression (Figure 1). The mask branch is a small FCN applied to each RoI, predicting a segmentation mask in a pixel-to-pixel manner. Mask R-CNN is simple to implement and train given the Faster R-CNN framework, which facilitates a wide range of flexible architecture designs. Additionally, the mask branch only adds a small computational overhead, enabling a fast system and rapid experimentation.

In principle Mask R-CNN is an intuitive extension of Faster R-CNN, yet constructing the mask branch properly is critical for good results. Most importantly, Faster R-CNN was not designed for pixel-to-pixel alignment between network inputs and outputs. This is most evident in how *RoIPool* [18, 12], the *de facto* core operation for attending to instances, performs coarse spatial quantization for feature extraction. To fix the misalignment, we propose a simple, quantization-free layer, called *RoIAlign*, that faithfully preserves exact spatial locations. Despite being

[1] Following common terminology, we use *object detection* to denote detection via *bounding boxes*, not masks, and *semantic segmentation* to denote per-pixel classification without differentiating instances. Yet we note that *instance segmentation* is both semantic and a form of detection.