# Mask Scoring R-CNN
# 一种为Mask打分的新方法

CVPR 2019 Oral 文章解读

冯芮苇
21821245
21821245@zju.edu.cn

## 摘要（Abstract）

*让深层网络意识到其预测质量是一个有趣但重要的问题。在实例分割任务中，大多数实例分割框架使用实例分类的置信度作为Mask的质量分数。然而，量化为Mask实例与其ground truth之间的IoU，这一结果通常与分类分数没有很强的相关性。本文研究了这个问题并提出了一个包含网络块的Mask Score R-CNN来学习预测实例Mask的质量。所提出的网络块将实例特征与对应的预测掩码组合以回溯Mask IoU。Mask scoring策略可以校正Mask质量和Mask Scoring之间的偏差，并通过在COCO AP评估过程中优先处理更准确的预测，从而提高分割性能。通过对COCO数据集的广泛评估，Mask Scoring R-CNN为不同的模型带来了一致和显著的好处，并且优于目前最先进的Mask R-CNN。我们希望我们简单有效的方法将为改进实例分割提供新的方向。*

## 1. Introduction

本文是一篇研究实例分割问题的文章，发表在2019年CVPR并获得了oral的机会。文章中提出的方法在coco图像实例分割任务上超越了何恺明的Mask R-CNN。文中提出了一种新的实例分割打分的方法，相比于Mask R-CNN的打分方法（根据目标区域的分类置信度打分），文中提出的打分方法能够实现和Mask质量具有更高的一致性。

### 1.1. Related work

目前，关于实例分割的方法大体上分为两类，一类是基于检测的分割方法，比如：Faster-RCNN、R-FCN等，首先获取每个实例的区域，然后在区域内预测每个实例的Mask；DeepMask方法通过对一个滑动窗口中心目标进行分割和分类从而实现实例分割；Instance-sensitive FCN、position-sensitive FCIS等通过生成位置敏感的映射，从而得到Mask；Mask R-CNN等通过增加语义分割分支，以及MaskLab等利用位置敏感分数来获得更好的结果。另一类方法是基于分割的实例分割方法：首先对每个像素的类别标签进行预测，然后将其分组形成实例分割结果。这类方法包括聚类、边缘检测、预测像素级的能量值等，以及一些其他的使用度量学习来学习嵌入的方法。然而，这两类方法在Mask的质量评价上存在较大的问题。通过分类置信度来衡量Mask的质量，会存在较大的偏差。如图1所示。分类置信度只用于区分proposal的语义类别，对于实例Mask的实际质量和完整性等并不能有更精确的表示。


图1 Mask R-CNN分数较高但是实际效果并不好的例子

### 1.2. Main contribution

本文的主要贡献如下：

提出了Mask Score R-CNN。这是第一个解决实例分割假设评分问题的框架，为提高实例分割模型的性能开辟了新的方向。文中的方法考虑到实例Mask的完整性，如果实例Mask的分类置信度较高但是预测所得的Mask不够好，则可以对实例Mask的分数进行扣分。

提出一种简单有效的分支MaskIoU head。在coco测试集上的实验结果表明当使用Mask Scoring R-CNN提出的Mask Score而不仅仅使用置信度时，AP字不同的主干网络上能够平均提高 1.5%。

## 2. Method

为了定量分析当前Mask R-CNN中存在的分类分数和Mask实际质量不完全匹配的问题，文中将Mask R-CNN的Mask Score与预测的Mask及其相应的Ground Truth Mask（MaskIoU）之间的实际IoU进行比较。

对比：Mask R-CNN（ResNet-18 FPN）coco2017 val-dataset

对比二者的关系（挑选detection proposal中二者都大于 0.5 的Soft-NMS检测假设）Mask R-CNN表现。

因此，文中提出了为每一个proposal都能根据MaskIoU得到一个校准的mask score。只是在Mask R-CNN的基础上加了一个MaskIoU Head。

MaskIoU在分类分数上的分布情况如图 2 (a)所示，每个MaskIoU区间的平均分类分数如图 2 (c)所示为蓝色，说明在Mask R-CNN中，分类分数与MaskIoU的相关性并不好。

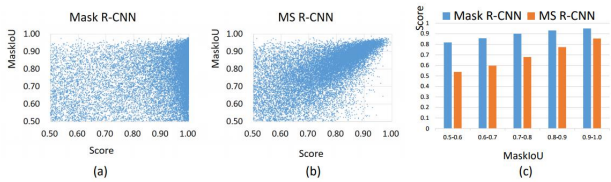


图 2 Mask R-CNN和MS R-CNN中Scoring

和MaskIOU相关性的对比

在大多数实例分割评估协议中，如COCO，检测假设的MaskIoU低，分数高是有害的。在许多实际应用中，确定检测结果何时可信以及何时不可用非常重要。这促使我们根据MaskIoU为每一个检测假设学习一个校准过的Mask Score 。在不丧失通用性的情况下，我们研究了Mask R-CNN框架，并提出了Mask Score R-CNN (MS R-CNN)，这是一个带有附加MaskIoU模块的MaskR-CNN，该模块学习Mask对齐的Mask评分。

### 2.1. Mask scoring R-CNN

Mask scoring R-CNN概念简单：Mask RCNN带MaskIoU Head，将实例特征与预测Mask一起作为输入，预测输入Mask与gtMask之间的IoU，如图 3 所示。

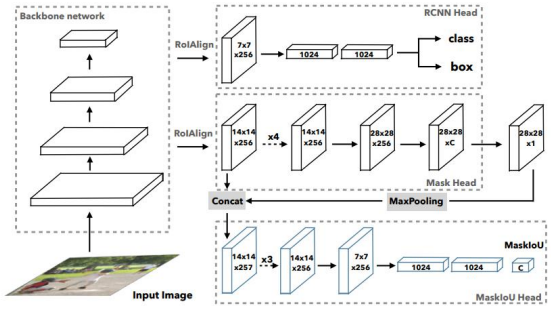


图 3 Mask RCNN with MaskIoU Head

### 2.2. Mask scoring

将Smask定义为预测Mask的评分。理想的smask等于预测mask和gt之间的像素级 IoU（之前表示为MaskIoU）；理想的smask只在gt类上有值而其他为 0。

Smask predicted Mask的最终得分

Scls 对proposal进行分类

Siou 侧重于回归MaskIOU

Smask = Scls • Siou

Mask score 对于语义类别信息和实例Mask的完整性都能够很好的体现。

### 2.3. MaskIOU head

在Mask R-CNN的基础上增加的module（MaskIOU head），让模型学会针对mask的得分规则。接收mask head的输出和ROI的feature作为输入，做一个简单的回归损失进行训练。只做gt class 的regression（test predicted）4 个卷积层 3 个全连接层。暗送秋波结构如图 2 所示。

### 2.4. Training

对于MaskIoU head的训练，我们使用RPN proposal作为训练样本。训练样本在proposal box和匹配的ground truth box之间需要有一个IoU大于 0.5，与Mask R-CNN的Mask head的训练样本相同。为了生成每个

训练样本的回归目标，我们首先得到目标类的预测Mask，并使用 0.5 的阈值对预测Mask进行二值化，然后利用二进制Mask与其匹配的ground truth之间的Mask作为目标Mask。我们使用l2损失来回归MaskIoU，损失权重设置为 1。将提出的MaskIoU head集成到Mask R-CNN中，对整个网络进行端到端训练。

### 2.5. Inference

在推理过程中，我们只使用MaskIoU head对R-CNN生成的分类分数进行校正。具体假设Mask R-CNN的R-CNN阶段输出N个边框，其中选取SoftNMS[2]后的top-k(即k = 100)计分框。然后将top-k盒输入MaskIoU head，生成多类Mask。这是标准的Mask R-CNN推断过程。我们也遵循这个过程，并输入top-k目标Mask来预测Mask。将预测的Mask与分类分数相乘，得到新的校准Mask分数作为最终Mask置信度。

## 3. Experiments

所有实验均在COCO数据集上进行，对象类别 80 个。我们遵循COCO 2017 设置，使用 115k图像训练分割进行训练，5k验证分割进行验证，20k测试开发分割进行测试。我们使用COCO评估指标AP(平均超过IoU阈值)报告结果，包括AP@0.5、AP@0.75 和APS、APM、APL(不同规模的AP)。AP@0.5(或AP@0.75)表示使用IoU阈值 0.5(或 0.75)来确定在评估中预测的边框或Mask是否为正。除非特别说明，AP使用mask IoU进行评估。首先设置了对比backbone和framework的实验，不同的backbone设置不同大小的输入（resize）。表格 1 是使用文中提出来的完整网络（mask scoring r-cnn）不同的backbone。APm：instance segmentation results；Apb：detection results.*（换backbone效果也很稳定）。

Table 1. COCO 2017 validation results. We report both detection and instance segmentation results. $AP_m$ denotes instance segmentation results and $AP_b$ denotes detection results. The results without ✓ are those of Mask R-CNN, while with ✓ are those of our MS R-CNN. The results show that our method is insensitive to different backbone networks.

| Backbone | MaskIoU head | $AP_m$ | $AP_m$@0.5 | $AP_m$@0.75 | $AP_b$ | $AP_b$@0.5 | $AP_b$@0.75 |
|---|---|---|---|---|---|---|---|
| ResNet-18 FPN | | 27.7 | 46.9 | 29.0 | 31.2 | 50.4 | 33.2 |
| | ✓ | 29.3 | 46.9 | 31.3 | 31.5 | 50.8 | 33.5 |
| ResNet-50 FPN | | 34.5 | 55.8 | 36.7 | 38.6 | 59.2 | 42.5 |
| | ✓ | 36.0 | 55.8 | 38.8 | 38.6 | 59.2 | 42.5 |
| ResNet-101 FPN | | 36.6 | 58.6 | 39.0 | 41.3 | 61.7 | 45.9 |
| | ✓ | 38.2 | 58.4 | 41.5 | 41.4 | 61.8 | 46.3 |

表格 2 是使用resnet101 为backbone换不同的网络。前两行是用的faster r-cnn，三四行用了FPN，最后两行用了DCN+FPN（RPN proposal）。

Table 2. COCO 2017 validation results. We report detection and instance segmentation results. $AP_m$ denotes instance segmentation results and $AP_b$ denotes detection results. In the results area, rows 1&2 use the Faster R-CNN framework; rows 3&4 additionally use FPN framework; rows 5&6 additionally use the DCN+FPN. The results show that consistent improvement of the proposed MaskIoU head.

| Backbone | MaskIoU head | FPN | DCN | $AP_m$ | $AP_m$@0.5 | $AP_m$@0.75 | $AP_b$ | $AP_b$@0.5 | $AP_b$@0.75 |
|---|---|---|---|---|---|---|---|---|---|
| ResNet-101 | | | | 33.9 | 53.9 | 36.2 | 38.6 | 57.3 | 42.8 |
| | ✓ | | | 35.0 | 54.0 | 37.7 | 38.7 | 57.4 | 43.0 |
| | | ✓ | | 36.6 | 58.6 | 39.0 | 41.3 | 61.7 | 45.9 |
| | ✓ | ✓ | | 38.2 | 58.4 | 41.5 | 41.4 | 61.8 | 46.3 |
| | | ✓ | ✓ | 37.7 | 60.3 | 40.0 | 42.9 | 63.4 | 47.8 |
| | ✓ | ✓ | ✓ | 39.1 | 60.0 | 42.4 | 43.1 | 63.5 | 47.7 |

试验结果表明：对于不同的backbone、不同的framework都是有明显的而且稳定的提升。

Table 3. Comparing different instance segmentation methods on COCO 2017 test-dev.

| Method | Backbone | AP | AP@0.5 | AP@0.75 | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|---|---|
| MNC [7] | ResNet-101 | 24.6 | 44.3 | 24.8 | 4.7 | 25.9 | 43.6 |
| FCIS [23] | ResNet-101 | 29.2 | 49.5 | - | - | - | - |
| FCIS+++ [23] | ResNet-101 | 33.6 | 54.5 | - | - | - | - |
| Mask R-CNN [15] | ResNet-101 | 33.1 | 54.9 | 34.8 | 12.1 | 35.6 | 51.1 |
| Mask R-CNN [15] | ResNet-101 FPN | 35.7 | 58.0 | 37.8 | 15.5 | 38.1 | 52.4 |
| Mask R-CNN [15] | ResNeXt-101 FPN | 37.1 | 60.0 | 39.4 | 16.9 | 39.9 | 53.5 |
| MaskLab [3] | ResNet-101 | 35.4 | 57.4 | 37.4 | 16.9 | 38.3 | 49.2 |
| MaskLab+ [3] | ResNet-101 | 37.3 | 59.8 | 36.6 | 19.1 | 40.5 | 50.6 |
| MaskLab+ [3] | ResNet-101 (JET) | 38.1 | 61.1 | 40.4 | 19.6 | 41.6 | 51.4 |
| Mask R-CNN | ResNet-101 | 34.3 | 55.0 | 36.6 | 13.2 | 36.4 | 52.2 |
| MS R-CNN | | 35.4 | 54.9 | 38.1 | 13.7 | 37.6 | 53.3 |
| Mask R-CNN | ResNet-101 FPN | 37.0 | 59.2 | 39.5 | 17.1 | 39.3 | 52.9 |
| MS R-CNN | | 38.3 | 58.8 | 41.5 | 17.8 | 40.4 | 54.4 |
| Mask R-CNN | ResNet-101 DCN+FPN | 38.4 | 61.2 | 41.2 | 18.0 | 40.5 | 55.2 |
| MS R-CNN | | 39.6 | 60.7 | 43.1 | 18.8 | 41.5 | 56.2 |

此外，文章还设计了丰富的消融实验。

1）研究MaskIOU head的不同输入设计对于最终结果的影响。具体方案如图所示：

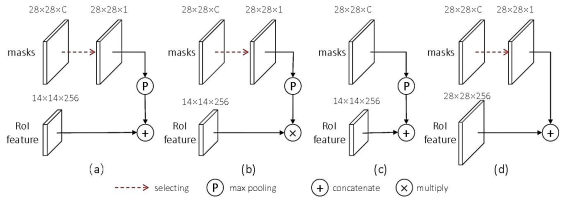


Figure 4. Different design choices of the MaskIoU head input.

实验结果表明，MaskIoU head对于不同种类的融合Mask预测和RoI特征的方法是鲁棒的。结果如表 4 所示：

表 4：不同融合方法对应的预测结果

| Setting | AP | AP@0.5 | AP@0.75 |
|---|---|---|---|
| Mask R-CNN baseline | 27.7 | 46.9 | 29.0 |
| (a) Target mask + RoI | 29.3 | 46.9 | 31.3 |
| (b) Target mask × RoI | 29.1 | 46.6 | 30.9 |
| (c) All masks + RoI | 29.1 | 46.6 | 30.8 |
| (d) Target mask + HR RoI | 29.1 | 46.7 | 31.1 |

2）文中设计的计算Mask Score是通过两个部分：一是mask classification，另一个是maskIoU regression。能否直接得到mask score呢？应该同时学习所有类别的MaskIoU么？因此，文中针对这一问题做了消融实验。调整不同的训练目标对比试验结果，如表 5 所示：

表 5：使用不同训练目标的训练结果

| Setting | AP | AP@0.5 | AP@0.75 |
|---|---|---|---|
| Mask R-CNN baseline | 27.7 | 46.9 | 29.0 |
| Setting #1: Target ins. | 29.3 | 46.9 | 31.3 |
| Setting #2: All cls. | 24.5 | 41.6 | 25.6 |
| Setting #3: Positive ins. | 28.2 | 45.5 | 30.2 |

3）使用不同的训练样本。从试验结果可以看出，使用全部的训练样本能够达到最好的效果（前提：boxlevel的IoU>0.5）。

表 6： 使用不同训练样本训练MaskIOU head结果对比

| Threshold | AP | AP@0.5 | AP@0.75 |
|---|---|---|---|
| $\tau = 0.0$ | 29.3 | 46.9 | 31.3 |
| $\tau = 0.3$ | 29.2 | 46.6 | 31.1 |
| $\tau = 0.5$ | 29.0 | 46.5 | 30.9 |
| $\tau = 0.7$ | 28.8 | 46.9 | 30.5 |

4）对比两个不同的backbone的效果，用gt和predicted mask-IoU的相关性来衡量预测的MaskIoU的质量。如图 5 所示。

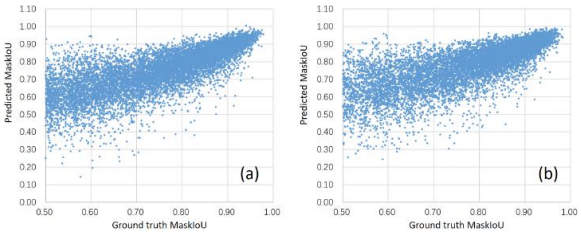


图 5：MaskIoU质量的可视化

5）设计实验给出MS R-CNN的表现上限。

表 7：对比MS R-CNN和Mask R-CNN换用不同的backbone的实验结果，以及对应的上限。

| Method | Backbone | AP |
|---|---|---|
| Mask R-CNN | ResNet-18 FPN | 27.7 |
| MS R-CNN | | 29.3 |
| MS R-CNN* | | 31.5 |
| Mask R-CNN | ResNet-101 DCN+FPN | 37.7 |
| MS R-CNN | | 39.1 |
| MS R-CNN* | | 41.7 |

## 4. Conclusion

本文研究了实例分割中为Mask打分的问题，通过在Mask R-CNN中添加一个MaskIoU head，实现Mask Score与MaskIoU的对齐（这在大多数实例分割框架中通常被忽略）。实验证明，文中提出的MaskIOU head非常有效，并且容易实施，能很好地进行迁移并具有良好的鲁棒性。它也可以应用于其他实例分割网络，以获得更可靠的Mask score。

## References


[1] Huang Z, Huang L, Gong Y, et al. Mask Scoring R-CNN[J]. 2019.

[2] M. Bai and R. Urtasun. Deep watershed transform for instance segmentation. In IEEE Conference on Computer Vision and Pattern Recognition, pages 2858–2866, 2017. 2

[3] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis. Softnmsimproving object detection with one line of code. In IEEE International Conference on Computer Vision, pages 5562–5570, 2017.

[4] L.-C. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang, and H. Adam. Masklab: Instance segmentation by refining object detection with semantic and direction features. arXiv preprint arXiv:1712.04837, 2017.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on Pattern Analysis and Machine Intelligence, pages 834–848, 2018.

[6] B. Cheng, Y. Wei, H. Shi, R. Feris, J. Xiong, and T. Huang. Revisiting rcnn: On awakening the classification power of faster rcnn. In European Conference on Computer Vision, pages 473–490, 2018.

[7] J. Dai, K. He, Y. Li, S. Ren, and J. Sun. Instance-sensitive fully convolutional networks. In European Conference on Computer Vision, pages 534–549, 2016.

[8] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In IEEE Conference on Computer Vision and Pattern Recognition, pages 3150– 3158, 2016.

[9] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In Advances in Neural Information Processing Systems, pages 379–387, 2016.

[10] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In IEEE International Conference on Computer Vision, pages 764–773, 2017.

[11] B. De Brabandere, D. Neven, and L. Van Gool. Semantic instance segmentation with a discriminative loss function. arXiv preprint arXiv:1708.02551, 2017.

[12] A. Fathi, Z. Wojna, V. Rathod, P. Wang, H. O. Song, S. Guadarrama, and K. P. Murphy. Semantic instance segmentation via deep metric learning. arXiv preprint arXiv:1703.10277, 2017.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In IEEE Conference on Computer Vision and Pattern Recognition, pages 580–587, 2014.