# CenterMask : Real-Time Anchor-Free Instance Segmentation

# CenterMask：实时无锚实例分割

Youngwan Lee and Jongyoul Park\*

李永万 和 朴钟宇\*

Electronics and Telecommunications Research Institute (ETRI), South Korea

电子与通信研究院（ETRI），韩国

{yw.lee, jongyou1 }@etri.re.kr

{yw.lee, jongyou1 }@etri.re.kr

# Abstract

# 摘要

We propose a simple yet efficient anchor-free instance segmentation, called CenterMask, that adds a novel spatial attention-guided mask (SAG-Mask) branch to anchor-free one stage object detector (FCOS [33]) in the same vein with Mask R-CNN [9]. Plugged into the FCOS object detector, the SAG-Mask branch predicts a segmentation mask on each detected box with the spatial attention map that helps to focus on informative pixels and suppress noise. We also present an improved backbone networks, VoVNetV2, with two effective strategies: (1) residual connection for alleviating the optimization problem of larger VoVNet [19] and (2) effective Squeeze-Excitation (eSE) dealing with the channel information loss problem of original SE. With SAG-Mask and VoVNetV2, we deign CenterMask and CenterMask-Lite that are targeted each to large and small models, respectively. Using the same ResNet-101-FPN backbone, Cen-terMask achieves 38.3%, surpassing all previous state-of-the-art methods while at a much faster speed. CenterMask-Lite also outperforms the state-of-the-art by large margins at over 35fps on Titan Xp. We hope that CenterMask and VoVNetV2 can serve as a solid baseline of real-time instance segmentation and backbone network for various vision tasks, respectively. The Code is available at https: //github.com/youngwanLEE/CenterMask.

我们提出了一种简单而高效的无需锚点实例分割方法，名为CenterMask，它在无锚单阶段目标检测器（FCOS [33]）的基础上增加了一个新颖的空间注意力引导掩码（SAG-Mask）分支，与Mask R-CNN [9]一脉相承。将SAG-Mask分支插入到FCOS目标检测器中，它会在每个检测到的框上预测一个分割掩码，并使用空间注意力图帮助聚焦于信息丰富的像素并抑制噪声。我们还提出了一个改进的骨干网络，VoVNetV2，包括两种有效策略：（1）残差连接，用于缓解较大VoVNet [19] 的优化问题；（2）有效的挤压-激励（eSE），处理原始SE的通道信息丢失问题。利用SAG-Mask和VoVNetV2，我们设计了CenterMask和CenterMask-Lite，分别针对大型和小型模型。使用相同的ResNet-101-FPN骨干网络，CenterMask达到了38.3%的性能，超过了所有之前的最先进方法，且速度更快。CenterMask-Lite在Titan Xp上以超过35fps的速度也大幅超越了现有最佳性能。我们希望CenterMask和VoVNetV2能够分别作为实时实例分割和各类视觉任务的可靠基线。

# 1. Introduction

# 1. 引言

Recently, instance segmentation has made great progress beyond object detection. The most representative method, Mask R-CNN [9], extended on object detection (e.g., Faster R-CNN [30]), has dominated COCO [23] benchmarks since instance segmentation can be easily solved by detecting objects and then predicting pixels on each box. However, even if there have been many works for improving the Mask R-CNN [9], few works exist for considering the speed of the instance segmentation. Although YOLACT [1] is the first real-time one-stage instance

最近，实例分割在物体检测的基础上取得了重大进展。最具代表性的方法，Mask R-CNN [9]，是在物体检测（例如，Faster R-CNN [30]）的基础上扩展的，自实例分割可以简单地通过检测物体然后在每个框上预测像素来实现以来，它一直主导着COCO [23]基准测试。然而，尽管已经有很多工作 旨在改进Mask R-CNN [9]，但很少有工作考虑到实例分割的速度。尽管YOLACT [1] 是第一个实时的一阶段实例

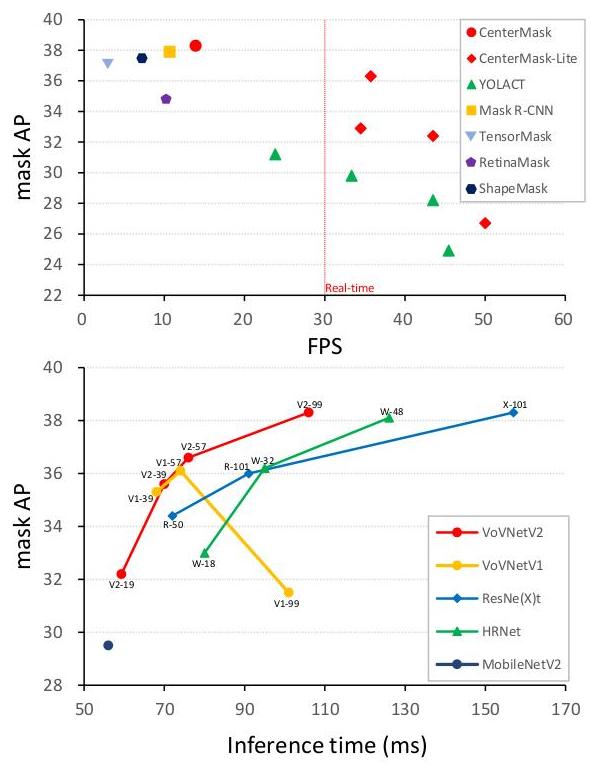


Figure 1: Accuracy-speed Tradeoff. across various instance segmentation models (top) and backbone networks (bottom) on COCO. The inference speed of Cen-terMask & CenterMask-Lite is reported on the same GPU (V100/Xp) with their counterparts. Note that all backbone networks in the bottom are compared under the proposed CenterMask. Please refer to section 3.2, Table 3 and Table 5 for details.

图1：准确度-速度权衡。在COCO上的各种实例分割模型（顶部）和基础网络（底部）之间的权衡。CenterMask & CenterMask-Lite的推理速度是在与它们的对比GPU（V100/Xp）上报告的。请注意，底部的所有基础网络都是在提出的CenterMask下进行比较的。具体细节请参考3.2节、表3和表5。

segmentation due to its parallel structure and extremely lightweight assembly process, the accuracy gap from Mask R-CNN [9] is still significant. Thus, we aim to bridge the gap by improving both accuracy and speed.

由于其并行结构和极其轻量级的组装过程，segmentation的准确度与Mask R-CNN [9]相比仍有显著差距。因此，我们旨在通过提高准确度和速度来缩小这一差距。

While Mask R-CNN [9] is based on a two-stage object detector (e.g., Faster R-CNN) that first generates box proposals and then predicts box location and classification, YOLACT [1] is built on one-stage detector (RetinaNet [22])

Mask R-CNN [9]基于两阶段物体检测器（例如，Faster R-CNN），首先生成框提议，然后预测框位置和分类，而YOLACT [1]是基于单阶段检测器（RetinaNet [22]）

\*Corresponding author.

\*通讯作者。

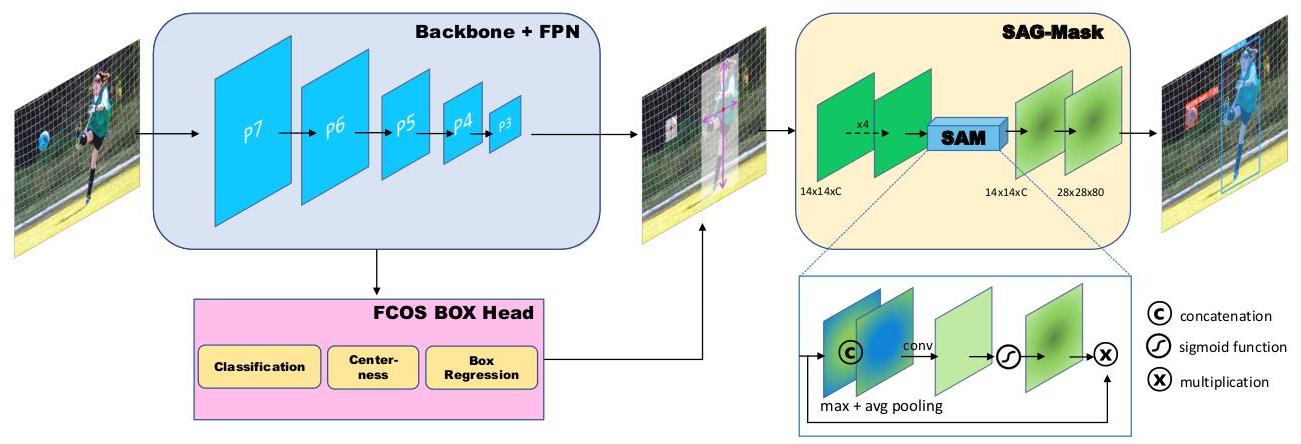


Figure 2: Architecture of CenterMask. where P3 (stride of ) to P7 (stride of ) denote the feature map in feature pyramid of backbone network. Using the features from the backbone, FCOS [33] predicts bounding boxes. Spatial Attention-Guided Mask (SAG-Mask) predicts segmentation mask inside of the each detected box with Spaital Attention Module (SAM) helping to focus on the informative pixels but also suppress the noise.

图2：CenterMask的架构。P3（步幅为 ）到P7（步幅为 ）表示主干网络特征金字塔中的特征图。利用来自主干网络的特征，FCOS [33] 预测边界框。空间注意力引导掩码（SAG-Mask）在每个检测到的框内预测分割掩码，空间注意力模块（SAM）帮助关注信息像素并抑制噪声。

that directly predicts boxes without proposal step. However, these object detectors rely heavily on pre-define anchors, which are sensitive to hyper-parameters (e.g., input size, aspect ratio, scales, etc.) and different datasets. Besides, since they densely place anchor boxes for higher recall rate, the excessively many anchor boxes cause the imbalance of positive/negative samples and higher computation/memory cost. To cope with these drawbacks of anchor boxes, recently, many works tend to escape from the anchor boxes toward anchor-free by using corner/center points, which leads to more computation-efficient and better performance compared to anchor box based detectors.

直接预测框而无需提议步骤。然而，这些目标检测器严重依赖预定义的锚点，这些锚点对超参数（例如，输入大小、纵横比、尺度等）和不同的数据集非常敏感。此外，由于它们密集地放置锚框以提高召回率，过多的锚框导致正/负样本的不平衡以及更高的计算/内存成本。为了应对锚框的这些缺点，最近，许多工作 倾向于通过使用角点/中心点逃离锚框，走向无锚点，这相比基于锚框的检测器带来了更高的计算效率和更好的性能。

Therefore, we design a simple yet efficient anchor-free one stage instance segmentation called CenterMask that adds a novel spatial attention-guided mask branch to the more efficient one-stage anchor-free object detector (FCOS [33]) in the same way with Mask R-CNN [9]. Figure 2 shows the overview of our CenterMask. Plugged into the FCOS [33] object detector, our spatial attention-guided mask (SAG-Mask) branch takes the predicted boxes from the FCOS [33] detector to predict segmentation masks on each Region of Interest (RoI). The spatial attention module (SAM) in the SAG-Mask helps the mask branch to focus on meaningful pixels and suppressing uninformative ones.

因此，我们设计了一种简单而高效的无锚点单阶段实例分割方法，称为CenterMask，它在更高效的单阶段无锚点目标检测器（FCOS [33]）中添加了一种新颖的空间注意力引导掩码分支，方式与Mask R-CNN [9]相同。图2展示了我们CenterMask的概述。插入到FCOS [33]目标检测器中，我们的空间注意力引导掩码（SAG-Mask）分支从FCOS [33]检测器获取预测框，以在每个兴趣区域（RoI）上预测分割掩码。SAG-Mask中的空间注意力模块（SAM）帮助掩码分支关注有意义的像素并抑制无信息的像素。

When extracting features on each RoI for mask prediction, each RoI pooling should be assigned considering the RoI scales. Mask R-CNN [9] proposes a new assignment function, called RoIAlign, that does not consider the input scale. Thus, we design a scale-adaptive RoI assignment function that considers the input scale and is a more suitable one-stage object detector. We also propose a more effective backbone network VoVNetV2 based on VoVNet [19] that shows better performance and faster speed than ResNet [10] and DenseNet [14] due to its One-shot Aggregation (OSA). In Figure 1 (bottom), We found that stacking the OSA modules in VoVNet makes the performance degradation (e.g., VoVNetV1-99). We see this phenomenon as the motivation of ResNet [10] because the backpropagation of gradient is disturbed. Thus, we add the residual connection [10] into each OSA module to ease optimization, which makes the VoVNet deeper and in turn, boosts the performance.

在对每个 RoI 进行掩码预测时提取特征时，应考虑 RoI 尺度来为每个 RoI 池化分配。Mask R-CNN [9] 提出了一种新的分配函数，称为 RoIAlign，它不考虑输入尺度。因此，我们设计了一种考虑输入尺度的尺度自适应 RoI 分配函数，这是一种更适合的一阶段目标检测器。我们还提出了一种基于 VoVNet [19] 的更有效的骨干网络 VoVNetV2，由于它的单次聚合（OSA）特性，其性能和速度均优于 ResNet [10] 和 DenseNet [14]。在图 1（底部）中，我们发现 VoVNet 中堆叠 OSA 模块会导致性能下降（例如 VoVNetV1-99）。我们认为这种现象是 ResNet [10] 的动机，因为梯度的反向传播受到了干扰。因此，我们在每个 OSA 模块中添加了残差连接 [10] 以简化优化，这使得 VoVNet 变得更深，进而提升了性能。

In the Squeeze-Excitation (SE) [13] channel attention module, it was found that the fully connected layers reduce the channel size, thereby reducing computational burden and unexpectedly causing channel information loss. Thus, we re-design the SE module as effective SE (eSE) replacing the two FC layers with one FC layer maintaining channel dimension, which prevents the information loss and in turn, improves the performance. With residual connection and eSE modules, We propose VoVNetV2 on various scales; from lightweight VoVNetV2-19, base VoVNetV2- 39/57 and large model VoVNetV2-99 that are correspond with MobileNet-V2 [11], ResNet-50/101 [10] & HRNet-W18/32 [32], and ResNeXt-32x8d [36].

在 Squeeze-Excitation (SE) [13] 的通道注意力模块中，发现全连接层减少了通道大小，从而降低了计算负担，但意外地导致了通道信息丢失。因此，我们将 SE 模块重新设计为有效的 SE（eSE），用一个保持通道维度的全连接层替换两个 FC 层，这防止了信息丢失，反过来提高了性能。带有残差连接和 eSE 模块，我们提出了适用于各种尺度的 VoVNetV2；从轻量级的 VoVNetV2-19，基础型 VoVNetV2-39/57，到大型模型 VoVNetV2-99，分别对应 MobileNet-V2 [11]、ResNet-50/101 [10] & HRNet-W18/32 [32]，以及 ResNeXt-32x8d [36]。

With SAG-Mask and VoVNetV2, we design Center-Mask and CenterMask-Lite that are targeted each to large and small models, respectively. The Extensive experiments demonstrate the effectiveness of CenterMask & CenterMask-Lite and VoVNetV2. Using the same ResNet- 101 backbone [10], CenterMask outperforms all previous state-of-the-art single models on the COCO [23] instance and detection tasks while at a much faster speed. CenterMask-Lite with VoVNetV2-39 bakcbone also achieves mask AP / box AP, outperforming the state-of-the-art real-time instance segmentation YOLACT [1] by gain, respectively, at over 35fps on Titan Xp.

使用 SAG-Mask 和 VoVNetV2，我们设计了 Center-Mask 和 CenterMask-Lite，分别针对大型和小型模型。大量实验证明了 CenterMask 和 CenterMask-Lite 以及 VoVNetV2 的有效性。使用相同的 ResNet-101 主干网络 [10]，CenterMask 在 COCO [23] 实例分割和检测任务上超过了所有之前的最先进单一模型，并且在速度上更快。CenterMask-Lite 配合 VoVNetV2-39 主干网络也达到了 掩码 AP / 框 AP，分别超过了最先进实时实例分割方法 YOLACT [1] 的提升，在 Titan Xp 上超过 35fps。

# 2. CenterMask

# 2. CenterMask

In this section, first, we review the anchor-free object detector, FCOS [33], which is a fundamental object detection part of our CenterMask. Next, we demonstrate the architecture of the CenterMask and describe how the proposed spatial attention-guided mask branch (SAG-Mask) is designed to plug into the FCOS [33] detector. Finally, a more effective backbone network, VoVNetV2, is proposed to boost the performance of CenterMask in terms of accuracy and speed.

在这一节中，首先，我们回顾了无锚点目标检测器 FCOS [33]，这是我们 CenterMask 的基本目标检测部分。接下来，我们展示了 CenterMask 的架构，并描述了如何设计所提出的空间注意力引导的掩码分支（SAG-Mask）以插入到 FCOS [33] 检测器中。最后，我们提出了一个更有效的主干网络 VoVNetV2，以提高 CenterMask 在准确性和速度方面的性能。

# 2.1. FCOS

# 2.1. FCOS

FCOS [33] is an anchor-free and proposal-free object detection in a per-pixel prediction manner as like FCN [26]. Almost state-of-the-art object detectors such as Faster R-CNN [30], YOLO [29], and RetinaNet [22] use the concept of the pre-defined anchor box which needs elaborate parameter tunning and complex calculation associated with box IoU in training. Without the anchor-box, the FCOS [33] directly predicts a 4D vector plus a class label at each spatial location on a level of feature maps. As shown in Figure 2, the 4D vector embeds the relative offsets from the four sides of a bounding box to the location (e.g., left, right, top and bottom). In addition, FCOS [33] introduces the centerness branch to predict the deviation of a pixel to the center of its corresponding bounding box, which improves the detection performance. Avoiding complex computation of anchor-boxes, FCOS [33] reduces memory/computation cost but also outperforms the anchor box based object detectors. Because of the efficiency and good performance of the FCOS [33], we design the proposed CenterMask built upon the FCOS [33] object detector.

FCOS [33] 是一种无需锚框和提案的无锚点目标检测方法，它以像素级预测的方式类似于 FCN [26]。几乎所有的最先进的目标检测器，如 Faster R-CNN [30]、YOLO [29] 和 RetinaNet [22]，都使用了预定义锚框的概念，这需要在训练过程中进行精细的参数调整和与框 IoU 相关的复杂计算。FCOS [33] 不使用锚框，而是直接在每个特征图级别的空间位置预测一个4D向量和一个类标签。如图2所示，4D向量包含了边界框四边相对于位置（例如，左、右、上和下）的相对偏移。此外，FCOS [33] 引入了中心性分支来预测像素相对于其对应边界框中心的偏差，这提高了检测性能。FCOS [33] 通过避免锚框的复杂计算，降低了内存/计算成本，同时优于基于锚框的目标检测器。由于 FCOS [33] 的高效性和良好性能，我们设计了基于 FCOS [33] 目标检测器的所提出 CenterMask。

# 2.2. Architecture

# 2.2. 架构

Figure 2 shows overall architecture of the CenterMask. CenterMask consists of three-part:(1) backbone for feature extraction, (2) FCOS [33] detection head, and (3) mask head. The procedure of masking objects is composed of detecting objects from the FCOS [33] box head and then predicting segmentation masks inside the cropped regions in a per-pixel manner.

图2展示了 CenterMask 的整体架构。CenterMask 由三部分组成：(1) 用于特征提取的主干网络，(2) FCOS [33] 检测头，以及(3) 掩码头。掩码对象的流程包括从 FCOS [33] 盒头检测对象，然后以像素级方式在裁剪区域内预测分割掩码。

# 2.3. Adaptive RoI Assignment Function

# 2.3. 自适应 RoI 分配函数

After object proposals are predicted in the FCOS [33] box head, CenterMask predicts segmentation masks using the predicted box regions in the same vein as Mask R-CNN. As the RoIs are predicted from different levels of feature maps in Feature Pyramid Network (FPN [21]), RoI Align [9] that extracts features should be assigned at different scales of feature maps with respect to RoI scales. Specifically, an RoI with a large scale has to be assigned to a higher feature level and vice versa. Mask R-CNN [9] based two-stage detector uses Equation 1 in FPN [21] to determine which feature map to be assigned.

在FCOS [33] 的框头中预测出对象建议后，CenterMask使用预测的框区域预测分割掩码，这与Mask R-CNN的方法类似。由于RoIs是从特征金字塔网络（FPN [21]）的不同层次的特征图中预测出来的，所以用于提取特征的RoI Align [9]应根据RoI的规模分配到不同尺度的特征图上。具体来说，一个较大规模的RoI必须分配到较高的特征层次，反之亦然。基于Mask R-CNN [9] 的两阶段检测器在FPN [21] 中使用公式1来确定要分配的特征图 。

where is 4 and are the width and height of the each RoI. However, Equation 1 is not suitable for CenterMask based one-stage detector because of two reasons. First, Equation 1 is tuned to two-stage detectors (e.g., FPN [21]) that use different feature levels compared to one-stage detectors (e.g, FCOS [33], RetinaNet [22]). Specifically, two-stage detectors use feature levels of P2 (stride of ) to P5 while one-stage detectors use from to that is larger receptive fields with lower-resolution. Besides, the canonical ImageNet pretraining size 224 in Equation 1 is hard-coded and not adaptive to feature scale variation. For example, when the input dimension is and the area of an RoI is , the RoI is assigned to relative higher feature P4 despite its small size of the area with respect to input dimension, which results in reducing small object AP. Therefore, we define Equation 2 as a new RoI assignment function suited for CenterMask based one-stage detectors.

其中 是4， 是每个RoI的宽度和高度。然而，公式1并不适用于基于CenterMask的一阶段检测器，原因有两个。首先，公式1是为两阶段检测器（例如，FPN [21]）调整的，这些检测器与一阶段检测器（例如，FCOS [33]、RetinaNet [22]）相比使用不同的特征层次。具体来说，两阶段检测器使用从P2（步长为 ）到P5 的特征层次，而一阶段检测器使用从 到 的特征层次，这具有更大的感受野和较低的分辨率。此外，公式1中的标准ImageNet预训练尺寸224是硬编码的，并且不适应特征尺度的变化。例如，当输入维度为 且RoI的面积为 时，尽管相对于输入维度RoI的面积较小，RoI仍被分配到相对较高的特征P4，这导致减少了小对象的AP。因此，我们定义了公式2作为新的RoI分配函数，适用于基于CenterMask的一阶段检测器。

where is the last level (e.g.,7) of feature map in backbone and are area of input image and the RoI, respectively. Without the canonical size 224 in Equation 1, Equation 2 adaptively assign RoI pooling scale by the ratio of input/RoI area. If is lower than minimum level (e.g., P3), is clamped to the minimum level. Specifically, if the area of an RoI is bigger than half of the input area, the RoI is assigned to the highest feature level(e.g., ). Inversely, while Equation 1 assigns to the RoI with , Equation 2 determine - 5 level which maybe minimum feature level for area of the RoI that is about smaller than input size. We can find that the proposed RoI assignment method improves the small object AP than Equation 1 because of its adaptive and scale-aware assignment strategy in Table 2. From an ablation study, we set to P5 and to P3.

其中 是骨干网络中特征图的最后一级（例如，7）， 分别是输入图像和RoI的面积。在公式1中没有224的标准大小，公式2通过输入/RoI面积的比例自适应地分配RoI池化尺度。如果 低于最小级别（例如，P3）， 将被限制在最小级别。具体来说，如果一个RoI的面积大于输入面积的一半，RoI将被分配到最高特征级别（例如， ）。相反，当公式1将 分配给具有 的RoI时，公式2确定 - 5 级，这可能是RoI面积约为 小于输入大小的最小特征级别。我们可以发现，在表2中提出的RoI分配方法由于其自适应和尺度感知的分配策略，提高了小对象的AP，超过了公式1。从消融研究中，我们设置 为P5， 为P3。

# 2.4. Spatial Attention-Guided Mask

# 2.4. 空间注意力引导掩码

Recently, attention methods [13, 34, 40, 28] have been widely applied to object detections because it helps to focus on important features, but also suppress unnecessary ones. In particular, channel attention [13, 12] emphasizes ’what’ to focus across channels of feature maps while spaital attention focuses ’where’ is an informative regions. Inspired by the spatial attention mechanism, we adopt a spatial attention module to guide the mask head for spotlighting meaningful pixels and repressing uninformative ones.

最近，注意力方法[13, 34, 40, 28]已被广泛应用于目标检测，因为它有助于关注重要特征，同时抑制不必要的特征。特别是，通道注意力[13, 12]强调在特征图的通道上“关注什么”，而空间注意力 关注“在哪里”是有信息的区域。受到空间注意力机制的启发，我们采用了一个空间注意力模块来引导掩码头，以突出有意义的像素并抑制不具信息的像素。

Thus, we design a spatial attention-guided mask (SAG-Mask), as shown in Figure 2. Once features inside the predicted RoIs are extracted by RoI Align [9] with resolution, those features are fed into four conv layers and

因此，我们设计了一个如图2所示的由空间注意力引导的掩码（SAG-Mask）。一旦通过RoI Align [9] 以 分辨率提取了预测RoIs内的特征，这些特征就会被送入四个卷积层中。

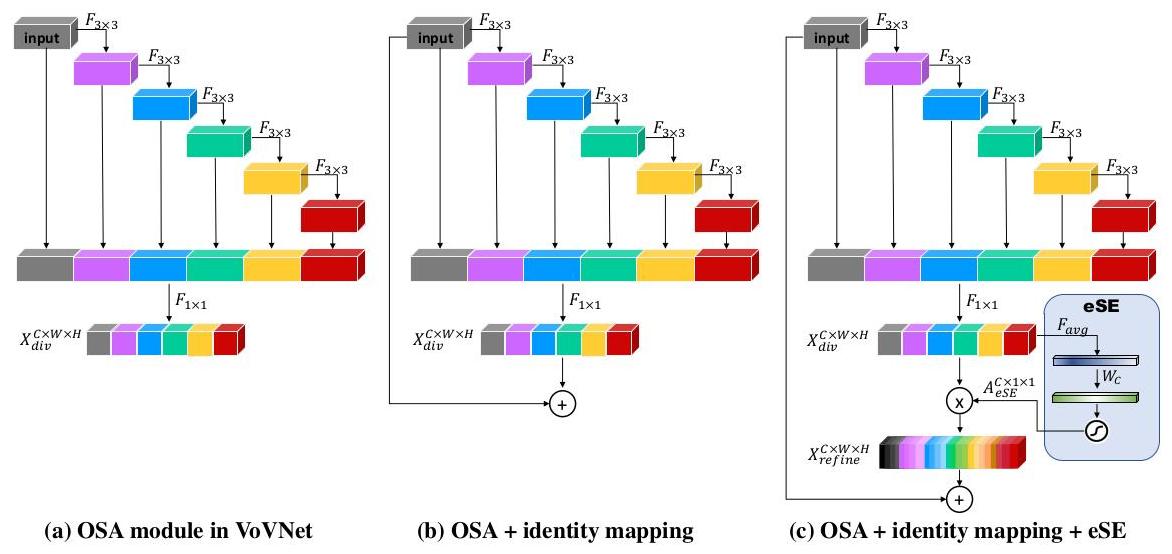


Figure 3: Comparison of OSA modules. denote conv layer respectively, is global average pooling, is fully-connected layer, is channel attention map, indicates element-wise multiplication and denotes element-wise addition.

图3：OSA模块的比较。 分别表示 卷积层， 是全局平均池化， 是全连接层， 是通道注意力图， 表示逐元素乘法， 表示逐元素加法。

spatial attention module (SAM) sequentially. To exploit the spatial attention map as a feature descriptor given input feature map , the SAM first generates pooled features by both average and max pooling operations respectively along the channel axis and aggregates them via concatenation. Then it is followed by a conv layer and normalized by the sigmoid function. The computation process is summarized as follow:

空间注意力模块（SAM）依次处理。为了利用空间注意力图 作为给定输入特征图 的特征描述符，SAM首先通过平均和最大池化操作分别在通道轴上生成池化特征 ，并通过连接将它们聚合。然后通过一个 卷积层，并由sigmoid函数归一化。计算过程总结如下：

where denotes the sigmoid function, is conv layer and represents concatenate operation. Finally, the attention guided feature map is computed as:

其中 表示sigmoid函数， 是 卷积层， 表示连接操作。最后，计算得到注意力引导的特征图 ：

where denotes element-wise multiplication. After then, a deconv upsamples the spatially attended feature map to resolution. Lastly, a conv is applied for predicting class-specific masks.

其中 表示逐元素乘法。之后，一个 反卷积上采样将空间注意特征图上采样到 分辨率。最后，应用一个 卷积来预测特定类别的掩码。

# 2.5. VoVNetV2 backbone

# 2.5. VoVNetV2 主干网络

In this section, we propose more effective backbone networks, VoVNetV2, for further boosting the performance of CenterMask. VoVNetV2 is improved from VoVNet [19] by adding residual connection [10] and the proposed effective Squeeze-and-Excitation (eSE) attention module to the VoVNet. VoVNet is a computation and energy efficient backbone network that can efficiently present diversified feature representation because of One-Shot Aggregation (OSA) modules. As shown in Figure 3(a) OSA module consists of consecutive conv layers and aggregates the subsequent feature maps at once, which can capture diverse receptive fields efficiently and in turn outperforms DenseNet and ResNet in terms of accuracy and speed.

在本节中，我们提出了更有效的骨干网络 VoVNetV2，以进一步提 CenterMask 的性能。VoVNetV2 是在 VoVNet [19] 的基础上改进的，通过添加残差连接 [10] 和我们提出的有效挤压和激发（eSE）注意力模块到 VoVNet 中。VoVNet 是一种计算和能量效率高的骨干网络，由于 One-Shot Aggregation（OSA）模块，能够有效地呈现多样化的特征表示。如图 3(a) 所示，OSA 模块由连续的卷积层组成，并一次性聚合后续的特征图，这可以有效地捕获多样化的感受野，进而比 DenseNet 和 ResNet 在准确性和速度上表现更佳。

Residual connection: Even with its efficient and diverse feature representation, VoVNet has a limitation in terms of optimization. As OSA modules are stacked (i.g., deeper) in VoVNet, we observe the accuracy of the deeper models is saturated or degradation. Specifically, Table 4 shows the accuracy of VoVNetV1-99 is lower than that of VoVNetV1- 57. Based on the motivation of ResNet [10], We conjecture that stacking OSA modules make the backpropagation of gradient gradually hard due to the increase of transformation functions such as conv. Therefore, as shown in Figure 3(b), we also add the identity mapping [10] to OSA modules. Correctly, the input path is connected to the end of an OSA module that is able to backpropagate the gradients of every OSA module in an end-to-end manner on each stage like ResNet. Boosting the performance of VoVNet, the identity mapping also makes the VoVNet possible to enlarge its depth such as VoVNet-99.

残差连接：尽管 VoVNet 具有高效且多样化的特征表示，但在优化方面存在局限性。随着 VoVNet 中 OSA 模块的堆叠（例如，更深），我们观察到更深模型的准确性饱和或退化。特别是，表 4 显示 VoVNetV1-99 的准确性低于 VoVNetV1-57。基于 ResNet [10] 的动机，我们推测堆叠 OSA 模块使得由于卷积等变换函数的增加，梯度反向传播逐渐变得困难。因此，如图 3(b) 所示，我们还在 OSA 模块中添加了恒等映射 [10]。正确地，输入路径连接到 OSA 模块的末端，能够以端到端的方式在每个阶段反向传播每个 OSA 模块的梯度，类似于 ResNet。提升 VoVNet 的性能，恒等映射也使得 VoVNet 能够增加其深度，如 VoVNet-99。

Effective Squeeze-Excitation (eSE): For further boosting the performance of VoVNet, We also propose a channel attention module, effective Squeeze-Excitation (eSE), improving original SE [13] more effectively. Squeeze-Excitation (SE) [13], a representative channel attention method adopted in CNN architectures, explicitly models the interdependency between the channels of feature maps to enhance its representation. The SE module squeezes the spatial dependency by global average pooling to learn a channel specific descriptor and then two fully-connected (FC) layers followed by a sigmoid function are used to rescale the input feature map to highlight only useful channels. In short, given input feature map , the channel attention map is computed as:

有效挤压-激发（eSE）：为了进一步提高VoVNet的性能，我们还提出了一个通道注意力模块，即有效挤压-激发（eSE），它能更有效地改进原始SE [13]。挤压-激发（SE）[13] 是一种在CNN架构中采用的代表性通道注意力方法，它显式地建模特征图通道之间的相互依赖性以增强其表示性。SE模块通过全局平均池化压缩空间依赖性，以学习特定于通道的描述符，然后通过两个全连接（FC）层后跟sigmoid函数来重新缩放输入特征图，以突出显示有用的通道。简而言之，给定输入特征图 ，通道注意力图 的计算如下：

where is channel-wise global average pooling, are weights of two fully-connected layers, denotes ReLU non-linear operator and indicates sigmoid function.

其中 是通道-wise全局平均池化， 是两个全连接层的权重， 表示ReLU非线性操作符， 表示sigmoid函数。

However, it is assumed that the SE module has a limitation: channel information loss due to dimension reduction. For avoiding high model complexity burden, two FC layers of the SE module need to reduce channel dimension. Specifically, while the first FC layer reduces input feature channels to using reduction ratio , the second FC layer expands the reduced channels to original channel size . As a result, this channel dimension reduction causes channel information loss.

然而，人们认为SE模块存在一个局限性：由于维度降低导致的通道信息丢失。为了避免高模型复杂度负担，SE模块的两个FC层需要降低通道维度。具体来说，第一个FC层使用降低比例 将输入特征通道 减少到 ，第二个FC层将降低的通道扩展到原始通道大小 。因此，这种通道维度降低导致了通道信息的丢失。

Therefore, we propose effective SE (eSE) that uses only one FC layer with channels instead of two FCs without channel dimension reduction, which rather maintains channel information and in turn improves performance. the eSE process is defined as:

因此，我们提出了有效SE（eSE），它只使用一个具有 通道的FC层，而不是两个FC层，且不进行通道维度降低，从而维护通道信息，进而提高性能。eSE过程定义为：

where is the diversified feature map computed by conv in OSA module. As a channel attentive feature descriptor, the is applied to the diversified feature map to make the diversified feature more informative. Finally, when using the residual connection, the input feature map is element-wise added to the refined feature map . The details of How the eSE module is plugged into the OSA module are shown in Figure 3(c).

其中 是由 OSA 模块中的 conv 计算得到的多样化特征图。作为一种通道注意力特征描述符， 被应用于多样化特征图 以使多样化特征更具信息性。最后，在使用残差连接时，输入特征图逐元素地添加到精炼的特征图 中。eSE 模块如何插入到 OSA 模块中的细节如图 3(c) 所示。

# 2.6. Implementation details

# 2.6. 实现细节

Since CenterMask is built on FCOS [33] object detector, we follow hyper-parameters of the FCOS [33] except for positive score threshold 0.03 instead of 0.05 Since FCOS [33] does not generate positive RoI samples well in initial training time. While using FPN levels 3 through 7 with 256 channels in the detection step, we use in the masking step, as mentioned in 2.3. We also use mask scoring [15] that recalibrates classification score considering predicted mask quality (e.g., mask IoU) in Mask R-CNN.

由于 CenterMask 是基于 FCOS [33] 目标检测器构建的，我们遵循 FCOS [33] 的超参数，除了将正分数阈值从 0.05 改为 0.03，因为 FCOS [33] 在初始训练阶段不能很好地生成正 RoI 样本。在检测步骤中，我们使用 FPN 层级的 3 到 7，每个层级有 256 个通道，在掩码步骤中，如 2.3 节所述，我们使用 。我们还使用了掩码评分 [15]，它在 Mask R-CNN 中通过考虑预测的掩码质量（例如，掩码 IoU）来重新校准分类分数。

CenterMask-Lite: To achieve real-time processing, we try to make the proposed CenterMask lightweight. We downsize three parts: backbone, box head, and mask head. In the backbone, first, we reduce the channels of FPN from 256 to 128, which can decrease the output of conv in FPN but also input dimension of box and mask head. And then, we replace the backbone network with more lightweight VoVNetV2-19 that has 4 OSA modules on each stage comprised of 3 conv layers instead of 5 as in VoVNetv2-39/57. In the box head, there are four conv layers with 256 channels on each classification and box branch where the centerness branch is shared with the box branch. We reduce the number of conv layer from 4 to 2 with 128 channels. Lastly, in the mask head, we also reduce the number of conv layers and channels in the feature extractor and mask scoring part from to , respectively.

CenterMask-Lite：为了实现实时处理，我们尝试使提出的CenterMask轻量化。我们缩小了三个部分：主干网络、框头和掩膜头。在主干网络中，首先，我们将FPN的通道数 从256减少到128，这样可以减少FPN中 卷积的输出，同时降低框头和掩膜头的输入维度。然后，我们将主干网络替换为更轻量级的VoVNetV2-19，该网络在每个阶段包含4个OSA模块，每个模块由3个卷积层组成，而不是VoVNetv2-39/57中的5个。在框头中，每个分类和框分支有四个 卷积层，每个层256个通道，其中中心性分支与框分支共享。我们将卷积层的数量从4个减少到2个，通道数为128。最后，在掩膜头中，我们还在特征提取器和掩膜评分部分分别减少了卷积层的数量和通道数 到 。

Training: We set the number of detection boxes from the FCOS [33] to 100, and the highest-scoring boxes are fed into the SAG-mask branch for training mask branch. We use the same mask target as Mask R-CNN that is made by the intersection between an RoI and its associated ground-truth mask. During training time, we define a multi-task loss on each RoI as:

训练：我们将FCOS[33]的检测框数量设置为100，并将得分最高的框输入到SAG掩膜分支以训练掩膜分支。我们使用与Mask R-CNN相同的掩膜目标，该目标是通过RoI与其关联的真实掩膜之间的交集生成的。在训练期间，我们为每个RoI定义了一个多任务损失：

where the classification loss , centerness loss , and box regression loss are same as those in [33] and is the average binary cross-entropy loss identical as in [9]. Unless specified, the input image is resized to have 800 pixels [21] along the shorter side and their longer side less or equal to 1333. We train CenterMask by using Stochastic Gradient Descent (SGD) for 90K iterations ( epoch) with a mini-batch of 16 images and initial learning rate of 0.01 which is decreased by a factor of 10 at and iterations, respectively. We use a weight decay of 0.0001 and a momentum of 0.9 , respectively. All backbone models are initialized by ImageNet pre-trained weights.

其中分类损失 、中心度损失 和框回归损失 与文献 [33] 中的相同， 是与文献 [9] 中相同的平均二元交叉熵损失。除非特别指定，输入图像的短边被调整到800像素 [21]，长边不超过1333像素。我们通过使用随机梯度下降（SGD）算法对CenterMask进行90K次迭代（ 个周期），每次迭代使用16张图像的迷你批次和初始学习率0.01，分别在 和 迭代时将学习率减少10倍。我们使用的权重衰减为0.0001，动量为0.9。所有基础模型均使用ImageNet预训练权重进行初始化。

Inference: At test time, the FCOS detection part yields 50 high-score detection boxes, and then the mask branch uses them to predict segmentation masks on each RoI. CenterMask/CenterMask-Lite use a single scale of 800/600 pixels for the shorter side, respectively.

推断：在测试时，FCOS检测部分产生50个高分检测框，然后掩膜分支使用这些检测框在每一个RoI上预测分割掩膜。CenterMask/CenterMask-Lite分别使用短边为800/600像素的单尺度。

# 3. Experiments

# 3. 实验

In this section, we evaluate the effectiveness of Center-Mask on COCO [23] benchmarks. All models are trained on the train 2017 and va12017 are used for ablation studies. Final results are reported on test-dev for comparison with state-of-the-arts. We use as mask average precision AP (averaged over IoU thresholds), , , and (AP at different scale). We also denote box as . All ablation studies are conducted using CenterMask with ResNet-50-FPN. We report the inference time of models using one thread (1 batch size) on the same enviroment equipped with Titan Xp GPU, CUDA v10.0,

在本节中，我们评估了Center-Mask在COCO [23]基准上的有效性。所有模型均在train 2017上训练，va12017用于消融研究。最终结果在test-dev上报告，以与现有技术水平进行比较。我们使用 作为掩码平均精度AP（在IoU阈值上平均）， 、 和 （不同尺度的AP）。我们还表示框 为 。所有消融研究均使用带有ResNet-50-FPN的CenterMask进行。我们报告了在配备Titan Xp GPU和CUDA v10.0的同一环境下，使用一个线程（1批大小）的模型推理时间，



Figure 4: Results of CenterMask with VoVNetV2-99 on COCO test-dev2017.

图4：CenterMask与VoVNetV2-99在COCO test-dev2017上的结果。

| Component | APmask | APbox | Time (ms) |
| --- | --- | --- | --- |
| FCOS (baseline), ours |  | 37.8 | 57 |
| + mask head (Eq. 1 [21]) | 33.4 | 38.3 | 67 |
| + mask head (Eq. 2, ours) | 33.8 | 38.7 | 67 |
| + SAM | 34.0 | 38.9 | 67 |
| + Mask scoring | 34.7 | 38.8 | 72 |

Table 1: Spatial Attention Guided Mask (SAG-Mask)

表1：空间注意力引导掩码（SAG-Mask）

These models use ResNet-50 backbone. We note that the mask heads with Eq. 1 is same as the mask branch of Mask R-CNN. SAM and Scoring denotes the proposed Spatial Attention Module and mask scoring [15].

这些模型使用ResNet-50作为主干网络。我们注意到，带有等式1的掩码头与Mask R-CNN的掩码分支相同。SAM和Scoring表示所提出的空间注意力模块和掩码评分[15]。

cuDNN v7.3, and pytorch1.1. The Qualitative results of CenterMask are shown in Figure 4.

cuDNN v7.3和pytorch1.1。CenterMask的定性结果在图4中显示。

# 3.1. Ablation study

# 3.1. 消融研究

Scale-adaptive RoI assignment function: Comparing to Equation 1 [21], we validate the proposed Equation 2 in CenterMask. Table 1 shows that our scale-adaptive RoI assignment function considering the input scale improves by over the counterpart. It means that Equation 2 regarding the ratio of input/RoI is more scale-adaptive than Equation 1.

尺度自适应RoI分配函数：与方程1 [21]相比，我们验证了CenterMask中的所提方程2。表1显示，考虑到输入尺度的我们的尺度自适应RoI分配函数比对比方法提高了 。这意味着，相对于方程1，关于输入/RoI比例的方程2更具尺度适应性。

Spatial Attention Guided Mask: Table 1 demonstrates the influence of each component in building Spatial Attention Guided Mask (SAG-Mask). The baseline, FCOS [33] object detector, starts from with the run time of . Adding only naive mask head improves the box performance by and obtains . With the prementioned scale-adaptive RoI mapping strategy, our spatial attention module, , makes the mask performance forward because the spatial attention module helps the mask predictor to focus on informative pixels but also suppress noise. It can also be seen that the detection performance is boosted when using SAM. We suggest that result from the SAM, the refined feature maps of mask head would also have a secondary effect on the detection branch that shares feature maps of the backbone.

空间注意力引导掩码：表1展示了构建空间注意力引导掩码（SAG-Mask）中每个组件的影响。基线，FCOS [33] 目标检测器，从 开始，运行时间为 。仅添加简单的掩码头，提高了框的性能 并获得 。通过预先提到的尺度自适应RoI映射策略，我们的空间注意力模块 ，使得掩码性能向前发展，因为空间注意力模块帮助掩码预测器关注信息丰富的像素，同时也抑制噪声。还可以看到，使用SAM时检测性能得到提升。我们推测，SAM的结果，细化后的掩码头特征图也会对共享主干特征图的检测分支产生二次影响。

| Feature Level | Apmask | APbox |
| --- | --- | --- |
|  | 34.4 | 38.8 |
| P3 P6 | 34.6 | 38.8 |
| P3 P5 | 34.6 | 38.9 |
| P3 P4 | 34.4 | 38.5 |

Table 2: Feature level ranges for RoIAlign [9] in CenterMmask. P3 P7 denotes the feature maps with output stride of

表2：CenterMmask中RoIAlign [9] 的特征级别范围。P3 P7表示输出步长为 的特征图

The SAG-mask also deploys the mask scoring [15] that recalibrates the score regarding the predicted mask IoU. As a result, the mask scoring increases performance by . We note that the mask scoring cannot boost detection performance because the recalibrated mask score adjusts the ranks of mask results in the evaluation step, not refines the features of the mask head like the SAM. Besides, SAM rarely causes extra computation while the mask scoring leads to computation overhead (e.g., ).

SAG-mask还部署了掩码评分 [15] ，该评分重新校准了关于预测掩码IoU的分数。因此，掩码评分通过 提高了性能。我们注意到掩码评分无法提升检测性能，因为重新校准的掩码分数在评估步骤中调整了掩码结果的排名，而不是像SAM那样细化掩码头的特征。此外，SAM很少引起额外的计算，而掩码评分导致了计算开销（例如， ）。

Feature selection. We also ablate which feature level range is suitable for our CenterMask based one-stage detector. Since FCOS [33] detector extract features from P3 P7, we start the same feature levels in the SAG-mask branch. As shown in Table 2, the performance of the P3 P7 range is not as good as other ranges. We speculate P7 feature map is too small to extract fine features for pixel-level prediction (e.g., ). We observe that feature range achieves the best result, which means feature maps with a bigger resolution are advantageous for the mask prediction.

特征选择。我们还研究了对于基于单阶段检测器CenterMask而言，哪个特征级别范围是合适的。由于FCOS [33] 检测器从 P3 P7 提取特征，我们在 SAG-mask 分支上从相同的特征级别开始。如表2所示，P3 P7 范围的性能不如其他范围。我们推测 P7 特征图太小，无法为像素级预测（例如， ）提取细微特征。我们观察到 特征范围取得了最佳效果，这意味着具有更高分辨率的特征图对于掩码预测更有利。

VoVNetV2: We extend VoVNet to VoVNetV2 by using residual connection and the proposed effective SE (eSE) module into the VoVNet. Table 4 shows residual connection consistently improves VoVNet-39/57/99. In particular, the reason that the improved AP margin of VoVNet-99 is bigger than VoVNet-39/57 is that VoVNet-99 comprised of more OSA modules can have more effect of residual connection that alleviates the optimization problem.

VoVNetV2：我们通过在 VoVNet 中使用残差连接和所提出的有效 SE（eSE）模块，将 VoVNet 扩展为 VoVNetV2。表4显示残差连接始终提高了 VoVNet-39/57/99 的性能。特别是，VoVNet-99 改进的 AP 边际之所以大于 VoVNet-39/57，是因为 VoVNet-99 由更多的 OSA 模块组成，可以更有效地利用残差连接来缓解优化问题。

| Backbone | Params. | Apmask |  |  |  | Apbox |  |  |  | Time (ms) |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| MobileNetV2 [31] | 28.7M | 29.5 | 12.0 | 31.4 | 43.8 | 32.6 | 17.8 | 35.2 | 43.2 | 56 |
| VoVNetV2-19 [19] | 37.6M | 32.2 | 14.1 | 34.8 | 48.1 | 35.9 | 20.8 | 39.2 | 47.6 | 59 |
| HRNetV2-W18 [32] | 36.4M | 33.0 | 14.3 | 34.7 | 49.9 | 36.7 | 20.7 | 39.4 | 49.3 | 80 |
| ResNet-50 [10] | 51.2M | 34.7 | 15.5 | 37.6 | 51.5 | 38.8 | 22.4 | 42.5 | 51.1 | 72 |
| VoVNetV1-39 [19] | 49.0M | 35.3 | 15.5 | 38.4 | 52.1 | 39.7 | 23.0 | 43.3 | 52.7 | 68 |
| VoVNetV2-39 | 52.6M | 35.6 | 16.0 | 38.6 | 52.8 | 40.0 | 23.4 | 43.7 | 53.9 | 70 |
| HRNetV2-W32 [32] | 56.2M | 36.2 | 16.0 | 38.4 | 53.0 | 40.6 | 23.0 | 43.8 | 53.1 | 95 |
| ResNet-101[10] | 70.1M | 36.0 | 16.5 | 39.2 | 54.4 | 40.7 | 23.4 | 44.3 | 54.7 | 91 |
| VoVNetV1-57 [19] | 63.0M | 36.1 | 16.2 | 39.2 | 54.0 | 40.8 | 23.7 | 44.2 | 55.3 | 74 |
| VoVNetV2-57 | 68.9M | 36.6 | 16.9 | 39.8 | 54.5 | 41.5 | 24.1 | 45.2 | 55.2 | 76 |
| HRNetV2-W48 [32] | 92.3M | 38.1 | 17.6 | 41.1 | 55.7 | 43.0 | 25.8 | 46.7 | 55.9 | 126 |
| ResNeXt-101 [36] | 114.3M | 38.3 | 18.4 | 41.6 | 55.4 | 43.1 | 26.1 | 46.8 | 55.7 | 157 |
| VoVNetV1-99 [19] | 83.6M | 31.5 | 13.5 | 33.5 | 46.5 | 35.3 | 19.7 | 38.1 | 46.6 | 101 |
| VoVNetV2-99 | 96.9M | 38.3 | 18.0 | 41.8 | 56.0 | 43.5 | 25.8 | 47.8 | 57.3 | 106 |

Table 3: CenterMask with other backbones on COCO val 2017. Note that all models are trained with a same manner (e.g., 12 epoch, 16 batch size, without train & test augmentation). The inference time is reported on same Titan Xp GPU.

表3：CenterMask 与其他基础模型在 COCO 验证集 2017 上的对比。注意，所有模型都以相同的方式进行训练（例如，12个周期，16个批量大小，没有训练和测试增强）。推理时间是在同一块 Titan Xp GPU 上报告的。

| Backbone | Params. | APmask | APbox | Time (ms) |
| --- | --- | --- | --- | --- |
| VoVNetV1-39 |  | 35.3 | 39.7 | 68 |
| + residual |  | 35.5 (+0.2) |  | 68 |
| + SE [13] |  | 34.6 (-0.7) | 39.0 (-0.7) | 70 |
| + eSE, ours | 52.6M | 35.6 (+0.3) |  | 70 |
| VoVNetV1-57 | 63.0M | 36.1 | 40.8 | 74 |
| + residual | 63.0M | 36.4 (+0.3) |  | 74 |
| + SE [13] | 65.9M | 35.9 (-0.2) | 40.8 | 77 |
| + eSE, ours | 68.9M |  | 41.5 (+0.7) | 76 |
| VoVNetV1-99 | 83.6M | 31.5 | 35.3 | 101 |
| + residual | 83.6M |  | 42.5 (+7.2) | 101 |
| + SE [13] | 88.0M | 37.1 (+5.6) | 41.9 (+6.6) | 107 |
| eSE, ours | 96.9M |  | 43.5 (+8.2) | 106 |

Table 4: VoVNetV2 Start from VoVNetV1, VoVNetV2 is improved by adding residual connection [10] and the proposed ef-fetive SE (eSE).

表4：VoVNetV2 从 VoVNetV1 开始，VoVNetV2 通过添加残差连接 [10] 和所提出的效果 SE（eSE）进行了改进。

To validate eSE, we also apply the SE [13] to the VoVNet and compare it with the proposed eSE. As shown in Table 4, the SE worsens the performance of VoVNet or has no effect because the diversified feature map of OSA module losses channel information due to channel dimension reduction in the SE. Contrary to the SE, our eSE maintaining channel information using only layer boosts both and from VoVNetV1 with slight computation.

为了验证 eSE，我们还把 SE [13] 应用到 VoVNet 上，并将其与所提出的 eSE 进行比较。如表 4 所示，SE 使得 VoVNet 的性能下降或者没有效果，因为 OSA 模块的多样化特征图在 SE 的通道维度降低时损失了通道信息。与 SE 相比，我们的 eSE 通过仅使用 层来保持通道信息，从而提高了 VoVNetV1 的 和 的性能，同时计算量略有增加。

Comparison to other backbones: We expand VoVNetV2 on various scales; large (V-99), base (V-39/57), and lightweight (V-19) which correspond to ResNeXt-32- 8d [36] & HRNet-W48 [32], ResNet-50/101 [10] & HRNet-W18/W32 [32], and MobileNetV2 [31], respectively. Table 3 and Figure 1 (bottom) demonstrate VoVNetV2 is well-balanced backbone network in terms of accuracy and speed. While VoVNetV1-39 already outperforms its counterparts, VoVNetV2-39 shows better performance than ResNet-50/HRNet-W18 by a large margin of at faster speeds, respectively. Especially, the gain of is bigger than , respectively. A similar result pattern is shown in VoVNetV2-57 with its counterparts.

与其他基础模型的比较：我们在不同的规模上扩展 VoVNetV2；大型（V-99）、基础型（V-39/57）和轻量型（V-19），分别对应 ResNeXt-32-8d [36] & HRNet-W48 [32]、ResNet-50/101 [10] & HRNet-W18/W32 [32] 和 MobileNetV2 [31]。表 3 和图 1（底部）表明 VoVNetV2 在准确性和速度方面是一个平衡良好的基础网络。虽然 VoVNetV1-39 已经超过了其对比模型，但 VoVNetV2-39 在更快速度下以 的较大优势显示出比 ResNet-50/HRNet-W18 更好的性能。特别是， 的增益大于 的增益。在 VoVNetV2-57 及其对比模型中，也显示了类似的结果模式。

For large model, showing much faster run time , VoVNetV2-99 achieves competitive or higher than ResNeXt-101-32x8d despite fewer model parameters. For small model, VoVNetV2-19 outperforms MobileNetV2 by a large margin of , with comparable speed.

对于大型模型，显示出更快的运行时间 ，VoVNetV2-99 在模型参数更少的情况下，达到了与 ResNeXt-101-32x8d 竞争性的 或更高的 。对于小型模型，VoVNetV2-19 以 的较大优势超过了 MobileNetV2，速度相当。

# 3.2. Comparison with state-of-the-arts methods

# 3.2. 与现有方法的比较

For further validation of the CenterMask, we compare the proposed CenterMask with state-of-the-art instance segmentation methods. As most methods use train augmentation, we also adopt the scale-jitter where the shorter image side is randomly sampled from [640, 800] pixels [8]. For Centermask-Lite, [580, 600] scale jittering is used for training. We train CenterMask and CenterMask-Lite for 24/36 epochs and 48 epochs, respectively. Note that we do not use test-time augmentation [8] (multi-scale). The other hyper-parameters are kept same as ablation study. For fair speed comparison, we inference models on the same GPU as counterparts. Specifically, since most large models are tested on V100 GPU and YOLACT [1] models are reported on Titan Xp GPU, we also report CenterMask models on V100 and CenterMask-Lite models on Xp.

为了进一步验证 CenterMask，我们将提出的 CenterMask 与最先进的实例分割方法进行了比较。由于大多数方法 使用训练增强，我们也采用了尺度抖动，其中图像的较短的边从 [640, 800] 像素中随机采样 [8]。对于 Centermask-Lite，训练时使用了 [580, 600] 尺度抖动。我们分别训练 CenterMask 和 CenterMask-Lite 24/36 个时期和 48 个时期。请注意，我们没有使用测试时间增强 [8]（多尺度）。其他超参数与消融研究保持相同。为了公平的速度比较，我们在与对照相同的 GPU 上推理模型。具体来说，由于大多数大型模型都在 V100 GPU 上进行测试，而 YOLACT [1] 模型是在 Titan Xp GPU 上报告的，我们还报告了在 V100 上的 CenterMask 模型和在 Xp 上的 CenterMask-Lite 模型。

| Method | Backbone | epochs | APmask |  |  |  | APbox |  |  |  | Time | FPS | GPU |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Mask R-CNN, ours | R-101-FPN | 24 | 37.9 | 18.1 | 40.3 | 53.3 | 42.2 | 24.9 | 45.2 | 52.7 | 94 | 10.6 | V100 |
| ShapeMask [17] | R-101-FPN | N/A | 37.4 | 16.1 | 40.1 | 53.8 | 42.0 | 24.3 | 45.2 | 53.1 | 125 | 8.0 | V100 |
| TensorMask [5] | R-101-FPN | 72 | 37.1 | 17.4 | 39.1 | 51.6 | - | - | - | - | 380 | 2.6 | V100 |
| RetinaMask [7] | R-101-FPN | 24 | 34.7 | 14.3 | 36.7 | 50.5 | 41.4 | 23.0 | 44.5 | 53.0 | 98 | 10.2 | V100 |
| CenterMask | R-101-FPN | 24 | 38.3 | 17.7 | 40.8 | 54.5 | 43.1 | 25.2 | 46.1 | 54.4 | 72 | 13.9 | V100 |
| CenterMask\* | R-101-FPN | 36 | 39.8 | 21.7 | 42.5 | 52.0 | 44.0 | 25.8 | 46.8 | 54.9 | 66 | 15.2 | V100 |
| Mask R-CNN, ours | X-101-FPN | 36 | 39.3 | 19.8 | 41.4 | 55.0 | 44.1 | 27.0 | 46.7 | 54.6 | 165 | 6.1 | V100 |
| CenterMask | X-101-FPN | 36 | 39.6 | 19.7 | 42.0 | 55.2 | 44.6 | 27.1 | 47.2 | 55.2 | 123 | 8.1 | V100 |
| CenterMask | V-99-FPN | 36 | 40.6 | 20.1 | 42.8 | 57.0 | 45.8 | 27.8 | 48.3 | 57.6 | 84 | 11.9 | V100 |
| CenterMask\* | V-99-FPN | 36 | 41.8 | 24.4 | 44.4 | 54.3 | 46.5 | 28.7 | 48.9 | 57.2 | 77 | 12.9 | V100 |
| YOLACT-400 [1] | R-101-FPN | 48 | 24.9 | 5.0 | 25.3 | 45.0 | 28.4 | 10.7 | 28.9 | 43.1 | 22 | 45.5 | Xp |
| CenterMask-Lite | M-v2-FPN | 48 | 26.7 | 9.0 | 27.0 | 40.9 | 30.2 | 14.2 | 31.9 | 40.9 | 20 | 50.0 | Xp |
| YOLACT-550 [1] | R-50-FPN | 48 | 28.2 | 9.2 | 29.3 | 44.8 | 30.3 | 14.0 | 31.2 | 43.0 | 23 | 43.5 | Xp |
| CenterMask-Lite | V-19-FPN | 48 | 32.4 | 13.6 | 33.8 | 47.2 | 35.9 | 19.6 | 38.0 | 45.9 | 23 | 43.5 | Xp |
| YOLACT-550 [1] | R-101-FPN | 48 | 29.8 | 9.9 | 31.3 | 47.7 | 31.0 | 14.4 | 31.8 | 43.7 | 30 | 33.3 | Xp |
| YOLACT-700 [1] | R-101-FPN | 48 | 31.2 | 12.1 | 33.3 | 47.1 | 33.7 | 16.8 | 35.6 | 45.7 | 42 | 23.8 | Xp |
| CenterMask-Lite | R-50-FPN | 48 | 32.9 | 12.9 | 34.7 | 48.7 | 36.7 | 18.7 | 39.4 | 48.2 | 29 | 34.5 | Xp |
| CenterMask-Lite | V-39-FPN | 48 | 36.3 | 15.6 | 38.1 | 53.1 | 40.7 | 22.4 | 43.2 | 53.5 | 28 | 35.7 | Xp |

Table 5: CenterMask instance segmentation and detection performance on COCO test-dev2017. Mask R-CNN, RetinaMask, and CenterMask are implemented on the same base code [27] and CenterMask\* is implemented on top of Detect ron [35]. R, X, V, and M denote ResNet, ResNeXt-32x8d, VoVNetV2, and MobileNetV2, respectively. For fair compariosn, these results are tested with one thread and single-scale.

表 5：CenterMask 在 COCO test-dev2017 上的实例分割和检测性能。Mask R-CNN、RetinaMask 和 CenterMask 都是在相同的基代码 [27] 上实现的，而 CenterMask\* 是在 Detect ron [35] 之上实现的。R、X、V 和 M 分别表示 ResNet、ResNeXt-32x8d、VoVNetV2 和 MobileNetV2。为了公平比较，这些结果都是使用一个线程和单尺度测试的。

Under the same ResNet-101 backbone, CenterMask outperforms all other counterparts in terms of both accuracy and speed. In particular, compared to RetinaMask [7] that has similar architecture (i.g., one-stage detector + mask branch), CenterMask achieves gain. In less than half training epochs, Cen-terMask also surpasses the dense sliding window method, TensorMask [5], by at faster speed. Furthermore, to the best of our knowledge, the CenterMask with VoVNetV2-99 is the first method to achieves at over 10 fps. It is noted that after first submission, Detectron2 [35] has been released that is a better baseline code. Thus, we also re-implement our CenterMask\* on top of Detect ron2 [35] and obtain further performance gain.

在相同的 ResNet-101 主干网络下，CenterMask 在准确度 和速度上均优于所有其他对比方法。特别是与具有类似架构（例如，单阶段检测器 + 掩码分支）的 RetinaMask [7] 相比，CenterMask 实现 的增益。在不到一半的训练周期内，CenterMask 也以 的速度超过了密集滑动窗口方法 TensorMask [5]。此外，据我们所知，使用 VoVNetV2-99 的 CenterMask 是第一个在超过 10 fps 的情况下实现 的方法。值得注意的是，在首次提交后，发布了更好的基线代码 Detectron2 [35]。因此，我们也在 Detectron2 [35] 的基础上重新实现了我们的 CenterMask\*，并获得了进一步的性能提升。

We also compare with YOLACT [1] that is the representative real-time instance segmentation. We use four kinds of backbones (e.g., MobileNetV2, VoVNetV2-19, VoVNetV2-39, and ResNet-50), which have a different accuracy-speed tradeoff. Table 5 and Figure 1 (top) demonstrate CenterMask-Lite is consistently superior to YOLACT in terms of accuracy and speed. Compared to YOLACT, all CenterMask-Lite models achieve over 30 fps speed with large margins of both and .

我们还与代表性的实时实例分割方法 YOLACT [1] 进行了比较。我们使用了四种不同的主干网络（例如，MobileNetV2、VoVNetV2-19、VoVNetV2-39 和 ResNet-50），它们具有不同的准确度-速度权衡。表5和图1（顶部）显示 CenterMask-Lite 在准确度和速度上始终优于 YOLACT。与 YOLACT 相比，所有 CenterMask-Lite 模型都以超过 30 fps 的速度实现了较大的 和 的提升。

# 4. Discussion

# 4. 讨论

# In Table 5, we observe that using the same ResNet-101 backbone, Mask R-CNN [9] shows better performance than

# 在表5中，我们观察到使用相同的 ResNet-101 主干网络，Mask R-CNN [9] 表现出比

CenterMask on small object. We conjecture that Mask R-CNN [9] uses larger feature maps (P2) than Center-Mask (P3) in which the mask branch can extract much finer spatial layout of an object than the P3 feature map. We note that there are still rooms for improving one-stage instance segmentation performance like techniques of Mask R-CNN [9].

CenterMask对小物体。我们推测Mask R-CNN [9]使用的特征图（P2）比Center-Mask（P3）更大，其中的掩膜分支能够比P3特征图提取到更精细的空间布局。我们注意到在一阶段实例分割性能提升方面仍有改进空间，例如Mask R-CNN [9]的技术 。

# 5. Conclusion

# 5. 结论

We have proposed a real-time anchor-free one-stage instance segmentation and more effective backbone networks. Adding spatial attention guided mask branch to the anchor-free one stage instance detection, CenterMask achieves state-of-the-art performance at real-time speed. The newly proposed VoVNetV2 backbone spanning from lightweight to larger models makes CenterMask well-balanced performance in terms of speed and accuracy. We hope Center-Mask will serve as a baseline for real-time instance segmentation. We also believe our proposed VoVNetV2 can be used as a strong and efficient backbone network for various vision tasks .

我们提出了一种实时的无锚点一阶段实例分割方法以及更有效的骨干网络。在无锚点一阶段实例检测中添加了空间注意力引导的掩膜分支，CenterMask在实时速度上达到了最先进的性能。新提出的VoVNetV2骨干网络从轻量级到大型模型，使得CenterMask在速度和精度上取得了平衡的性能。我们希望Center-Mask能成为实时实例分割的基线。我们也相信我们提出的VoVNetV2可以作为各种视觉任务的强大且高效的骨干网络 。

Acknowledgements. We thank Hyung-il Kim and the DeepView team in ETRI. This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. B0101-15-0266, Development of High Performance Visual BigData Discovery Platform for Large-Scale Realtime Data Analysis and No. 2020-0-00004, Development of Previsional Intelligence based on Long-term Visual Memory Network).

致谢。我们感谢Hyung-il Kim和ETRI的DeepView团队。这项工作得到了韩国政府（MSIT）资助的信息通信技术规划与评估研究所（IITP）的资助（编号B0101-15-0266，用于大规模实时数据分析的高性能视觉大数据发现平台开发，以及编号2020-0-00004，基于长期视觉记忆网络的预测智能开发）。

After the initial submission, Detectron2 [35] has been released and we has developed the improved CenterMask\* on top of the Detect ron2 [35].

在初始提交后，Detectron2 [35] 已经发布，我们在Detectron2 [35]的基础上开发了改进的CenterMask\*。

# References

# 参考文献

[1] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In ICCV, 2019. 1,

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In CVPR, pages 6154- 6162, 2018. 1, 8

[3] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaox-iao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In CVPR, pages 4974-4983, 2019. 1, 8

[4] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In CVPR, pages 5659-5667, 2017. 3

[5] Xinlei Chen, Ross Girshick, Kaiming He, and Piotr Dollar. Tensormask: A foundation for dense object segmentation. In ICCV, 2019. 7, 8

[6] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qing-ming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In ICCV, 2019. 2

[7] Cheng-Yang Fu, Mykhailo Shvets, and Alexander C Berg. Retinamask: Learning to predict masks improves state-of-the-art single-shot detection for free. arXiv preprint arXiv:1901.03353, 2019. 7, 8

[8] Kaiming He, Ross Girshick, and Piotr Dollar. Rethinking imagenet pre-training. In ICCV, 2019. 7

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Gir-shick. Mask r-cnn. In

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016. 2, 4, 7

[11] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco An-dreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017. 2

[12] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Andrea Vedaldi. Gather-excite: Exploiting feature context in convolutional neural networks. In NIPS, pages 9401-9411, 2018. 3

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In CVPR, pages 7132-7141, 2018. 2, 3, 4, 7

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In CVPR, 2017. 2

[15] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In CVPR, pages 6409-6418, 2019. 1, 5, 6

[16] Youngjoo Jo and Jongyoul Park. Sc-fegan: Face editing generative adversarial network with user’s sketch and color. In ICCV, pages 1745-1753, 2019. 8

[17] Weicheng Kuo, Anelia Angelova, Jitendra Malik, and Tsung-Yi Lin. Shapemask: Learning to segment novel objects by refining shape priors. In

[18] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In ECCV, pages 734-750, 2018. 2

[19] Youngwan Lee, Joong-won Hwang, Sangrok Lee, Yuseok Bae, and Jongyoul Park. An energy and gpu-computation efficient backbone network for real-time object detection. In CVPR Workshops, pages 0-0, 2019. 1, 2, 4, 7

[20] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. arXiv preprint arXiv:1901.01892, 2019. 1

[21] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. In CVPR, 2017. 3, 5, 6

[22] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. ICCV, 2017. 1, 3

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV, 2014. 1, 2, 5

[24] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In CVPR, pages 8759-8768, 2018. 1

[25] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In ECCV, 2016. 7

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In CVPR, pages 3431-3440, 2015. 3

[27] Francisco Massa and Ross Girshick. maskrenn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. https://github.com/facebookresearch/ maskrcnn-benchmark, 2018. 8

[28] Zheng Qin, Zeming Li, Zhaoning Zhang, Yiping Bao, Gang , Yuxing Peng, and Jian Sun. Thundernet: Towards real-time generic object detection on mobile devices. In ICCV, 2019. 3

[29] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In CVPR, 2016. 3

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In NIPS, 2015. 1, 3

[31] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zh-moginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In CVPR, pages 4510-4520, 2018. 7

[32] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514, 2019. 2, 7

[33] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In ICCV, 2019.

[34] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In ECCV, pages 3-19, 2018. 3

[35] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github. com/facebookresearch/detectron2,2019.8

[36] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In CVPR, 2017. 2, 7

[37] Kimin Yun, Yongjin Kwon, Sungchan Oh, Jinyoung Moon, and Jongyoul Park. Vision-based garbage dumping action detection for real-world surveillance platform. ETRI Journal, 41(4):494-505, 2019. 8

[38] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In arXiv preprint arXiv:1904.07850, 2019. 2

[39] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In CVPR, pages 850-859, 2019. 2

[40] Xizhou Zhu, Dazhi Cheng, Zheng Zhang, Stephen Lin, and Jifeng Dai. An empirical study of spatial attention mechanisms in deep networks. In