Bridging the Gap Between Anchor-based and Anchor-free Detection via Adaptive Training Sample Selection

Abstract

Object detection has been dominated by anchor-based detectors for several years. Recently, anchor-free detectors have become popular due to the proposal of FPN and Focal Loss. In this paper, we first point out that the essential difference between anchor-based and anchor-free detection is actually how to define positive and negative training samples, which leads to the performance gap between them. **If they adopt the same definition of positive and negative samples during training, there is no obvious difference in the final performance, no matter regressing from a box or a point.** This shows that how to select positive and negative training samples is important for current object detectors. Then, we propose an Adaptive Training Sample Selection (ATSS) to automatically select positive and negative samples according to **statistical characteristics of object**. It significantly improves the performance of anchor-based and anchor-free detectors and bridges the gap between them. Finally, we discuss the necessity of tiling multiple anchors per location on the image to detect objects. Extensive experiments conducted on MS COCO support our aforementioned analysis and conclusions. With the newly introduced ATSS, we improve state-of-the-art detectors by a large margin to 50.7% AP without introducing any overhead. The code is available at https://github.com/sfzhang15/ATSS.

几年来，物体检测一直被基于锚的检测器所主导。最近，由于FPN和Focal Loss的提出，无锚检测器开始流行。在本文中，我们首先指出，基于锚点的检测和无锚点的检测的本质区别实际上是如何定义正负训练样本，这导致了它们之间的性能差距。**如果它们在训练过程中采用相同的正负样本定义，那么无论从一个盒子还是一个点回归，最终的性能都没有明显的差别。这表明如何选择正负训练样本对于当前的物体检测器来说非常重要。然后，我们提出了一个自适应训练样本选择（ATSS），根据物体的统计特征自动选择正负样本。它极大地提高了基于锚的检测器和无锚检测器的性能，并弥补了它们之间的差距。最后，我们讨论了在图像上每个位置铺设多个锚点来检测物体的必要性。**在MS COCO上进行的广泛实验支持我们的上述分析和结论。通过新引入的ATSS，我们在不引入任何开销的情况下，将最先进的检测器大幅提高到50.7% AP。该代码可在https://github.com/sfzhang15/ATSS。

1. Introduction

Object detection is a long-standing topic in the field of computer vision, aiming to detect objects of predefined categories. Accurate object detection would have far reaching impact on various applications including image recognition and video surveillance. In recent years, with the development of convolutional neural network (CNN), object detection has been dominated by anchor-based detectors, which can be generally divided into one-stage methods [36, 33] and two-stage methods [47, 9]. Both of them first tile a large number of preset anchors on the image, then predict the category and refine the coordinates of these anchors by one or several times, finally output these refined anchors as detection results. Because two-stage methods refine anchors several times more than one-stage methods, the former one has more accurate results while the latter one has higher computational efficiency. State-of-the-art results on common detection benchmarks are still held by anchor-based detectors.

Recent academic attention has been geared toward anchor-free detectors due to the emergence of FPN [32] and Focal Loss [33]. Anchor-free detectors directly find objects without preset anchors in two different ways. **One way is to first locate several pre-defined or self-learned keypoints and then bound the spatial extent of objects. We call this type of anchor-free detectors as keypoint-based methods [26, 71]. Another way is to use the center point or region of objects to define positives and then predict the four distances from positives to the object boundary. We call this kind of anchor-free detectors as center-based methods [56, 23].** These anchor-free detectors are able to eliminate those hyper-parameters related to anchors and have achieved similar performance with anchor-based detectors, making them more potential in terms of generalization ability.

由于FPN[32]和Focal Loss[33]的出现，最近学术界对无锚检测器给予了关注。无锚检测器以两种不同的方式直接找到没有预设锚的物体。**一种方式是首先定位几个预设的或自我学习的关键点，然后约束物体的空间范围。我们把这种类型的无锚检测器称为基于关键点的方法[26, 71]。另一种方法是使用物体的中心点或区域来定义阳性点，然后预测阳性点到物体边界的四个距离。我们把这种无锚检测器称为基于中心的方法[56, 23]。**这些无锚检测器能够消除那些与锚有关的超参数，并取得了与基于锚的检测器相似的性能，使它们在泛化能力方面更有潜力。

Among these two types of anchor-free detectors, keypoint-based methods follow the standard keypoint estimation pipeline that is different from anchor-based detectors. However, center-based detectors are similar to anchor-based detectors, which treat points as preset samples instead of anchor boxes. Take the one-stage anchor-based detector RetinaNet [33] and the center-based anchor-free detector FCOS [56] as an example, **there are three main differences between them: (1) The number of anchors tiled per location. RetinaNet tiles several anchor boxes per location, while FCOS tiles one anchor point1 per location. (2) The definition of positive and negative samples. RetinaNet resorts to the Intersection over Union (IoU) for positives an negatives, while FCOS utilizes spatial and scale constraints to select samples. (3) The regression starting status. RetinaNet regresses the object bounding box from the preset anchor box, while FCOS locates the object from the anchor point.** As reported in [56], the anchor-free FCOS achieves much better performance than the anchor-based RetinaNet, it is worth studying which of these three differences are essential factors for the performance gap.

在这两类无锚检测器中，基于关键点的方法遵循标准的关键点估计管道，与基于锚的检测器不同。然而，基于中心的检测器与基于锚的检测器相似，它将点作为预设样本而不是锚盒。以单阶段基于锚的检测器RetinaNet[33]和基于中心的无锚检测器FCOS[56]为例，**它们之间有三个主要区别。(1) 每个位置的锚点的数量。RetinaNet为每个位置铺设多个锚点盒，而FCOS为每个位置铺设一个锚点1。(2) 正负样本的定义。RetinaNet采用 "联合交集"（IoU）来定义正负样本，而FCOS利用空间和尺度约束来选择样本。(3) 回归的起始状态。RetinaNet从预设的锚箱回归物体边界箱，而FCOS从锚点定位物体。**如文献[56]所述，无锚的FCOS取得了比基于锚的RetinaNet好得多的性能，值得研究的是这三个差异中哪些是造成性能差距的基本因素。

In this paper, we investigate the differences between anchor-based and anchor-free methods in a fair way by **strictly ruling out all the implementation inconsistencies** between them. It can be concluded from experiment results that the **essential difference between these two kind of methods is the definition of positive and negative training samples**, which results in the performance gap between them. **If they select the same positive and negative samples during training, there is no obvious gap in the final performance, no matter regressing from a box or a point.** Therefore, how to select positive and negative training samples deserves further study. Inspired by that, we propose a new Adaptive Training Sample Selection (ATSS) to **automatically select positive and negative samples based on object characteristics.** It bridges the gap between anchor-based and anchor-free detectors. Besides, **through a series of experiments, a conclusion can be drawn that tiling multiple anchors per location on the image to detect objects is not necessary.** Extensive experiments on the MS COCO [34] dataset support our analysis and conclusions. State-of-the-art AP 50.7% is achieved by applying the newly

introduced ATSS without introducing any overhead. The main contributions of this work can be summarized as:

• Indicating the essential difference between anchor-based and anchor-free detectors is actually how to define positive and negative training samples.

• Proposing an adaptive training sample selection to automatically select positive and negative training samples according to statistical characteristics of object.

• Demonstrating that tiling multiple anchors per location on the image to detect objects is a useless operation.

• Achieving state-of-the-art performance on MS COCO without introducing any additional overhead.

在本文中，我们通过**严格地排除它们之间所有的实施不一致，公平地研究了基于锚的方法和无锚的方法之间的差异。从实验结果中可以得出结论，这两种方法的**本质区别在于正负训练样本的定义，**这导致了它们之间的性能差距。**如果他们在训练中选择相同的正负样本，那么无论从一个盒子还是一个点回归，最终的性能都不会有明显的差距。受此启发，我们提出了一种新的自适应训练样本选择（ATSS），**根据物体特征自动选择正负样本。**它弥补了基于锚和无锚检测器之间的差距。此外，**通过一系列的实验，可以得出一个结论：在图像上每个位置铺设多个锚来检测物体是没有必要的。**在MS COCO[34]数据集上进行的广泛实验支持了我们的分析和结论。通过应用新引入的ATSS，在不引入任何开销的情况下实现了最先进的AP50.7%。这项工作的主要贡献可以总结为：。

指出基于锚的检测器和无锚检测器之间的本质区别实际上是如何定义积极和消极的训练样本。

提出了一种自适应的训练样本选择，根据物体的统计特征自动选择正负训练样本。

证明在图像上每个位置铺设多个锚点来检测物体是一种无用的操作。

在MS COCO上实现了最先进的性能，而没有引入任何额外的开销。

2. Related Work

Current CNN-based object detection consists of anchor-based and anchor-free detectors. The former one can be divided into two-stage and one-stage methods, while the latter one falls into keypoint-based and center-based methods.

目前基于CNN的物体检测包括基于锚的和无锚的检测器。前者可分为两阶段和单阶段方法，而后者则分为基于关键点和基于中心的方法。

2.1. Anchor-based Detector

Two-stage method.

The emergence of Faster R-CNN [47] establishes the dominant position of two-stage anchor-based detectors. Faster R-CNN consists of a separate region proposal network (RPN) and a region-wise prediction network (R-CNN) [14, 13] to detect objects. After that, lots of algorithms are proposed to improve its performance, including architecture redesign and reform [4, 9, 5, 28, 30], context and attention mechanism [2, 51, 38, 7, 44], multi-scale training and testing [54, 41], training strategy and loss function [40, 52, 61, 17], feature fusion and enhancement [25, 32], better proposal and balance [55, 43]. Nowadays, state-of-the-art results are still held by two-stage anchor-based methods on standard detection benchmarks.

Faster R-CNN[47]的出现确立了基于锚点的两阶段检测器的主导地位。Faster R-CNN由一个独立的区域建议网络（RPN）和一个区域预测网络（R-CNN）组成[14, 13]，用于检测物体。之后，提出了很多算法来提高其性能，包括架构的重新设计和改革[4, 9, 5, 28, 30]，上下文和注意力机制[2, 51, 38, 7, 44]，多尺度训练和测试[54, 41]，训练策略和损失函数[40, 52, 61, 17]，特征融合和增强[25, 32]，更好的提议

和平衡[55, 43]。现在，在标准检测基准上，最先进的结果仍然由基于两阶段锚的方法保持。

One-stage method.

With the advent of SSD [36], one-stage anchor-based detectors have attracted much attention because of their high computational efficiency. SSD spreads out anchor boxes on multi-scale layers within a ConvNet to directly predict object category and anchor box offsets. Thereafter, plenty of works are presented to boost its performance in different aspects, such as fusing context information from different layers [24, 12, 69], training from scratch [50, 73], introducing new loss function [33, 6], anchor refinement and matching [66, 67], architecture redesign [21, 22], feature enrichment and alignment [35, 68, 60, 42, 29]. At present, one-stage anchor-based methods can achieve very close performance with two-stage anchor-based methods at a faster inference speed.

随着SSD[36]的出现，基于锚点的单阶段检测器因其高计算效率而备受关注。SSD在ConvNet中的多尺度层上铺开锚箱，直接预测物体类别和锚箱偏移。此后，大量的工作被提出来以提高其在不同方面的性能，如融合不同层的上下文信息[24, 12, 69]，从头开始训练[50, 73]，引入新的损失函数[33, 6]，锚的细化和匹配[66, 67]，架构重新设计[21, 22]，特征丰富和对齐[35, 68, 60, 42, 29]。目前，基于一阶段锚的方法可以在较快的推理速度下达到与基于二阶段锚的方法非常接近的性能。

2.2. Anchor-free Detector

Keypoint-based method.

This type of anchor-free method first locates several pre-defined or self-learned keypoints, and then generates bounding boxes to detect objects. CornerNet [26] detects an object bounding box as a pair of keypoints (top-left corner and bottom-right corner) and **CornerNet-Lite [27] introduces CornerNet-Saccade and CornerNet-Squeeze to improve its speed.** The second stage of **Grid R-CNN [39] locates objects via predicting grid points with the position sensitive merits of FCN and then determining the bounding box guided by the grid.** ExtremeNet [71] detects four extreme points (top-most, left-most, bottom-most, right-most) and one center point to generate the object bounding box. Zhu et al. [70] use keypoint **estimation** to find center point of objects and **regress to all other properties including size, 3D location, orientation and pose**. CenterNet [11] extends CornetNet as a triplet rather than a pair of keypoints to improve both precision and recall. RepPoints [65] **represents objects as a set of sample points** and learns to arrange themselves **in a manner that bounds the spatial extent of an object** and indicates semantically significant local areas.

这种类型的无锚方法首先定位几个预先定义的或自学的关键点，然后生成边界框来检测物体。
CornerNet[26]将物体边界框检测为一对关键点（左上角和右下角），CornerNet-Lite[27]引入了CornerNet-Saccade和CornerNet-Squeeze来提高其速度。Grid R-CNN[39]的第二阶段通过预测具有FCN的位置敏感优点的网格点来定位物体，然后在网格的引导下确定边界框。ExtremeNet[71]检测四个极端点（最上、最左、最下、最右）和一个中心点来生成物体的边界盒。Zhu等人[70]使用关键点估计来寻找物体的中心点，并回归到所有其他属性，包括尺寸、三维位置、方向和姿势。CenterNet[11]将CornetNet扩展为一个三联体而不是一对关键点，以提高精度和召回率。RepPoints[65]将物体表示为一组样本点，并学习以一种限定物体空间范围的方式排列它们，并指出有语义意义的局部区域。

Center-based method.

This kind of anchor-free method regards the center (e.g., the center point or part) of object as foreground to define positives, and then predicts the distances from positives to the four sides of the object bounding box for detection. YOLO [45] divides the image into an S × S grid, and the **grid cell that contains the center of an object is responsible for detecting this object**. DenseBox [20] uses a filled circle located in the center of the object to define positives and then predicts the four distances from positives to the bound of the object bounding box for

location. GA-RPN [59] defines the pixels in the center region of the object as positives to predict the location, width and height of object proposals for Faster R-CNN. FSAF [72] attaches an anchor-free branch with online feature selection to RetinaNet. The newly added branch defines the center region of the object as positives to locate it via predicting four distances to its bounds. FCOS [56] regards all the locations inside the object bounding box as positives with four distances and a novel centerness score to detect objects. CSP [37] only defines the center point of the object box as positives to detect pedestrians with fixed aspect ratio. FoveaBox [23] regards the locations in the middle part of object as positives with four distances to perform detection.

这种无锚方法将物体的中心（如中心点或部分）视为前景来定义阳性，然后预测阳性到物体边界框的四边的距离进行检测。YOLO[45]将图像分为S×S网格，包含物体中心的网格单元负责检测这个物体。DenseBox[20]使用位于物体中心的填充圆来定义阳性体，然后预测阳性体到物体边界框的四个距离来定位。GA-RPN[59]将物体中心区域的像素定义为阳性，以预测Faster R-CNN的物体建议的位置、宽度和高度。FSAF[72]在RetinaNet上附加了一个具有在线特征选择的无锚分支。新增加的分支将物体的中心区域定义为正数，通过预测其边界的四个距离来定位物体。FCOS[56]将物体边界框内的所有位置视为阳性，用四个距离和一个新的 centerness score来检测物体。CSP[37]只将物体框的中心点定义为阳性，以检测具有固定长宽比的行人。FoveaBox[23]将物体中间部分的位置视为正数，用四个距离来进行检测。

2. Difference Analysis of Anchor-based and Anchor-free Detection

Without loss of generality, the **representative** anchor-based RetinaNet [33] and anchor-free FCOS [56] are adopted to **dissect their differences.** In this section, we focus on the last two differences: **the positive/negative sample definition and the regression starting status. The remaining one difference: the number of anchors tiled per location,** will be discussed in subsequent section. Thus, we just tile one square anchor per location for RetinaNet, which is quite similar to FCOS. In the remaining part, we first introduce the experiment settings, then rule out all the implementation inconsistencies, finally point out the essential difference between anchor-based and anchor-free detectors.

在不丧失一般性的情况下，采用具有代表性的基于锚的RetinaNet[33]和无锚的FCOS[56]来剖析它们的差异。在这一节中，我们重点讨论最后两个差异：正/负样本定义和回归起始状态。剩下的一个差异：每个位置的锚点数量，将在随后的章节中讨论。因此，对于RetinaNet，我们只为每个位置铺设一个方形锚，这与FCOS非常相似。在剩下的部分，我们首先介绍了实验设置，然后排除了所有的实现不一致的地方，最后指出了基于锚和无锚检测器之间的本质区别。

3.1. Experiment Setting

Dataset.

All experiments are conducted on the challenging MS COCO [34] dataset that includes 80 object classes. Following the common practice [33, 56], all 115K images in the trainval35k split is used for training, and all 5K images in the minival split is used as validation for analysis study. We also submit our main results to the evaluation server for the final performance on the test-dev split.

所有的实验都是在具有挑战性的MS COCO[34]数据集上进行的，其中包括80个物体类别。按照通常的做法[33, 56]，在trainval35k分割中的所有115K图像被用于训练，在minival分割中的所有5K图像被用作分析研究的验证。我们还将我们的主要结果提交给评估服务器，以便在测试-开发分割中获得最终的性能。

Training Detail.

We use the ImageNet [49] pretrained ResNet-50 [16] with 5-level feature pyramid structure as the backbone. The newly added layers are initialized in the same way as in [33]. For RetinaNet, each layer in the 5-level feature pyramid is associated with one square anchor with 8S scale, where S is the total stride size. During training, we resize the input images to keep their shorter side being 800 and their longer side less or equal to 1, 333. The whole network is trained using the Stochastic Gradient Descent (SGD) algorithm for 90K iterations with 0.9 momentum, 0.0001 weight decay and 16 batch size. We set the initial learning rate as 0.01 and decay it by 0.1 at iteration 60K and 80K, respectively. Unless otherwise stated, the afore-mentioned training details are used in the experiments.

我们使用ImageNet[49]预训练的ResNet-50[16]，以5级特征金字塔结构为骨架。新增加的层的初始化方式与[33]中的方式相同。对于RetinaNet，5级特征金字塔中的每一层都与一个8S比例的方形锚相关联，其中S是总跨度大小。在训练过程中，我们调整输入图像的大小，以保持其短边为800，长边小于或等于1，333。整个网络使用随机梯度下降（SGD）算法进行了90K次迭代，动量为0.9，权重衰减为0.0001，批次大小为16。我们设定初始学习率为0.01，并在迭代60K和80K时分别衰减0.1。除非另有说明，上述的训练细节在实验中都有使用。

Inference Detail.

During the inference phase, we resize the input image in the same way as in the training phase, and then forward it through the whole network to output the predicted bounding boxes with a predicted class. After that, we use the preset score 0.05 to filter out plenty of background bounding boxes, and then output the top 1000 detections per feature pyramid. Finally, the Non-Maximum Suppression (NMS) is applied with the IoU threshold 0.6 per class to generate final top 100 confident detections per image.

在推理阶段，我们以与训练阶段相同的方式调整输入图像的大小，然后通过整个网络转发，输出带有预测类别的预测边界框。之后，我们使用预设分数0.05来过滤掉大量的背景边界框，然后输出每个特征金字塔的前1000个检测结果。最后，非最大抑制（NMS）被应用于每个类别的IoU阈值0.6，以产生每个图像的最终前100个置信检测。

3.2. Inconsistency Removal

We mark the anchor-based detector RetinaNet with only one square anchor box per location as RetinaNet (#A=1), which is almost the same as the anchor-free detector FCOS. However, as reported in [56], FCOS outperforms RetinaNet (#A=1) by a large margin in AP performance on the MS COCO minival subset, i.e., 37.1% vs. 32.5%. Furthermore, some new improvements have been made for FCOS including **moving centerness to regression branch, using GIoU loss function and normalizing regression targets by corresponding strides.** These improvements boost the AP performance of FCOS from 37.1% to 37.8% 2 , making the gap even bigger. However, part of the AP gap between the anchor-based detector (32.5%) and the anchor-free detector (37.8%) results from some universal improvements that are proposed or used in FCOS, such as **adding GroupNorm [62] in heads**, **using the GIoU [48] regression loss function, limiting positive samples in the ground-truth box [56], introducing the centerness branch [56] and adding a trainable scalar [56] for each level feature pyramid.** These improvements can also be applied to anchor-based detectors, therefore they are not the essential differences between anchor-based and anchor-free methods. We apply them to RetinaNet (#A=1) one by one so as to **rule out these implementation inconsistencies.** As listed in Table 1, these **irrelevant differences** improve the anchor-based RetinaNet to 37.0%, which still has a gap of 0.8% to the anchor-free FCOS. By now, **after removing all the irrelevant differences, we can explore the essential differences between anchor-based and anchor-free detectors in a quite fair way.**

Table 1: Analysis of implementation inconsistencies between RetinaNet and FCOS on MS COCO `minival` set. "#A=1" means there is one square anchor box per location.

Inconsistency	FCOS	RetinaNet (#A=1)					
GroupNorm	✓		✓	✓	✓	✓	✓
GIoU Loss	✓			✓	✓	✓	✓
In GT Box	✓				✓	✓	✓
Centerness	✓					✓	✓
Scalar	✓						✓
AP (%)	37.8	32.5	33.4	34.9	35.3	36.8	37.0

我们将每个位置只有一个方形锚箱的基于锚的检测器RetinaNet标记为RetinaNet（#A=1），这与无锚检测器FCOS几乎相同。然而，正如文献[56]所报道的，FCOS在MS COCO minival子集上的AP性能比RetinaNet(#A=1)高出一大截，即37.1%比32.5%。此外，FCOS还做了一些新的改进，包括**将中心点移至回归分支，使用GIoU损失函数，并通过相应的步长将回归目标归一化。**这些改进使FCOS的AP性能从37.1%提高到37.8% 2，使差距变得更大。然而，基于锚的检测器（32.5%）和无锚检测器（37.8%）之间的部分AP差距来自于FCOS中提出或使用的一些通用改进，例如**在头部添加GroupNorm[62]，使用GIoU[48]回归损失函数，限制真实箱中的正样本[56]，引入中心度分支[56]，以及为每一级特征金字塔添加一个可训练标度[56]。**这些改进也可以应用于基于锚的检测器，**因此它们不是基于锚和无锚方法之间的本质区别。**我们将它们逐一应用于RetinaNet（#A=1），以排除这些实施上的不一致。如表1所示，这些不相关的差异将基于锚的RetinaNet提高到37.0%，与无锚的FCOS仍有0.8%的差距。至此，在去除所有不相关的差异后，我们可以以一种相当公平的方式探索基于锚和无锚检测器之间的本质差异。

3.3. Essential Difference

After applying those universal improvements, these are only two differences between the anchor-based RetinaNet (#A=1) and the anchor-free FCOS. **One is about the classification sub-task in detection, i.e., the way to define positive and negative samples. Another one is about the regression sub-task, i.e., the regression starting from an anchor box or an anchor point.**

在应用这些普遍的改进后，基于锚的RetinaNet（#A=1）和无锚的FCOS之间只有两个区别。**一个是关于检测中的分类子任务，即定义正面和负面样本的方式。另一个是关于回归子任务，即从锚箱或锚点开始的回归。**

Classification.

As shown in Figure 1(a), **RetinaNet utilizes IoU to divide the anchor boxes from different pyramid levels into positives and negatives. It first labels the best anchor box of each object and the anchor boxes with IoU > θp *as positives, then regards the anchor boxes with IoU* < θn as negatives, finally other anchor boxes are ignored during training.** As shown in Figure 1(b), **FCOS uses spatial and scale constraints to divide the anchor points from different pyramid levels. It first considers the anchor points within the ground-truth box as candidate positive samples, then selects the final positive samples from candidates based on the scale range defined for each pyramid level3 , finally those unselected anchor points are negative samples.** As shown in Figure 1, **FCOS first uses the spatial constraint to find candidate positives in the spatial dimension, then uses the scale constraint to select final positives in the scale dimension.** In contrast, **RetinaNet utilizes IoU to directly select the final positives in the spatial and scale dimension simultaneously.** These two different sample selection strategies produce different positive and negative samples. As listed in the first column of Table 2 for RetinaNet (#A=1), using the spatial and scale constraint strategy instead of the IoU strategy improves the AP performance from 37.0% to 37.8%. As for FCOS, if it uses the IoU

strategy to select positive samples, the AP performance decreases from 37.8% to 36.9% as listed in the second column of Table 2. These results demonstrate that the definition of positive and negative samples is an essential difference between anchor-based and anchor-free detectors.

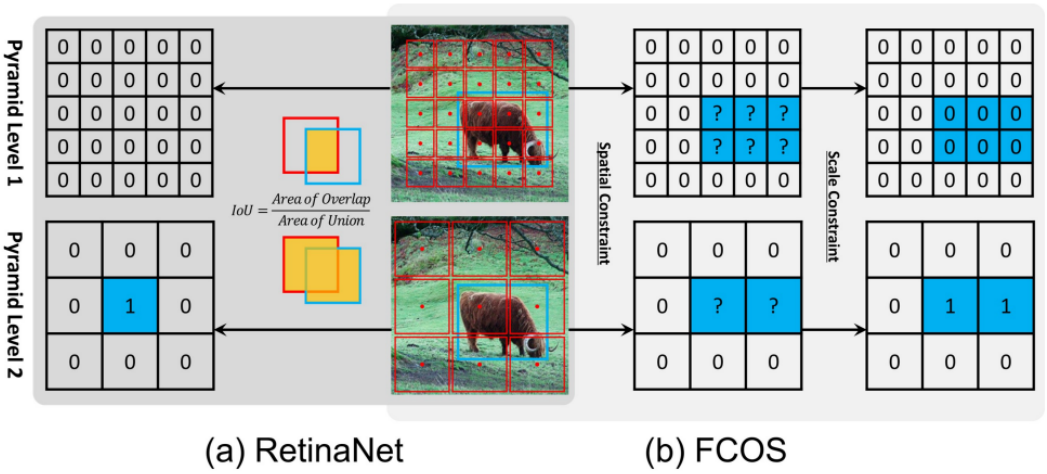


Figure 1: Definition of positives (1) and negatives (0). Blue box, red box and red point are ground-truth, anchor box and anchor point. (a) RetinaNet uses IoU to select positives (1) in spatial and scale dimension simultaneously. (b) FCOS first finds candidate positives (?) in spatial dimension, then selects final positives (1) in scale dimension.

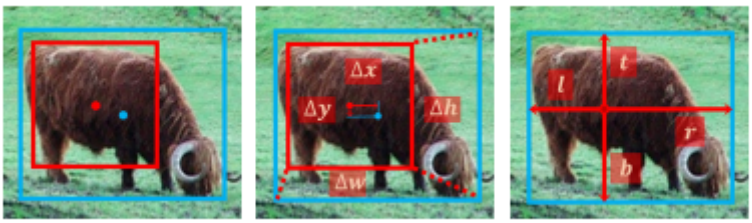
Table 2: Analysis of differences (%) between RetinaNet and FCOS on the MS COCO `minival` set.

Classification \ Regression	Box	Point
Intersection over Union	37.0	36.9
Spatial and Scale Constraint	37.8	37.8

如图1(a)所示，RetinaNet利用IoU将来自不同金字塔级别的锚点盒划分为阳性和阴性。**它首先将每个物体的最佳锚点盒和IoU>θ*p*的锚点盒标记为阳性，然后将*IoU*<*θ*n的锚点盒视为阴性，最后在训练中忽略其他锚点盒。如图1(b)所示，FCOS 利用空间和尺度约束来划分来自不同金字塔级别的锚点。它首先将地面实况框内的锚点视为候选正样本，然后根据为每个金字塔级别定义的尺度范围3从候选样本中选择最终的正样本，最后那些未选择的锚点为负样本。**如图1所示，FCOS**首先使用空间约束来寻找空间维度上的候选阳性样本，然后使用尺度约束来选择尺度维度上的最终阳性样本。**相反，RetinaNet利用IoU同时在空间和尺度维度上直接选择最终的阳性。这两种不同的样本选择策略产生了不同的阳性和阴性样本。如表2第一列所列的RetinaNet（#A=1），使用空间和尺度约束策略而不是IoU策略，将AP性能从37.0%提高到37.8%。至于FCOS，如果它使用IoU策略来选择阳性样本，则AP性能从37.8%下降到36.9%，如表2第二列所列。这些结果表明，阳性和阴性样本的定义是基于锚的检测器和无锚检测器之间的一个基本区别。

Regression.

After positive and negative samples are determined, the location of object is regressed from positive samples as shown in Figure 2(a). RetinaNet regresses from the anchor box with four offsets between the anchor box and the object box as shown in Figure 2(b), while FCOS regresses from the anchor point with four distances to the bound of object as shown in Figure 2(c). It means that for a positive sample, **the regression starting status of RetinaNet is a box while FCOS is a point**. However, as shown in the first and second rows of Table 2, when RetinaNet and FCOS adopt the same sample selection strategy to have consistent positive/negative samples, there is no obvious difference in final performance, no matter regressing starting from a point or a box, i.e., 37.0% vs. 36.9% and 37.8% vs. 37.8%. These results indicate that the regression starting status is an irrelevant difference rather than an essential difference. Conclusion. According to these experiments conducted in a fair way, we indicate that the essential difference between one-stage anchor-based detectors and center-based anchor-free detectors is actually **how to define positive and negative training samples, which is important for current object detection and deserves further study.**



(a) Positive sample (b) RetinaNet (c) FCOS

Figure 2: (a) Blue point and box are the center and bound of object, red point and box are the center and bound of anchor. (b) RetinaNet regresses from anchor box with four offsets. (c) FCOS regresses from anchor point with four distances.

在确定了正负样本后，从正样本中回归物体的位置，如图2（a）所示。如图2(b)所示，RetinaNet从锚箱回归，锚箱和物体箱之间有四个偏移，而FCOS从锚点回归，与物体的边界有四个距离，如图2(c)所示。这意味着对于阳性样本，**RetinaNet的回归起始状态是一个盒子，而FCOS是一个点。**然而，如表2第一行和第二行所示，当RetinaNet和FCOS采用相同的样本选择策略以获得一致的正/负样本时，无论从点还是从框开始回归，最终的性能都没有明显差异，即37.0%对36.9%，37.8%对37.8%。这些结果表明，回归起始状态是一个无关紧要的差异，而不是一个本质的差异。结论。根据这些以公平方式进行的实验，我们表明基于单阶段锚的检测器和基于中心的无锚检测器之间的本质区别实际上是**如何定义积极和消极的训练样本，这对当前的物体检测很重要，值得进一步研究。**

4. Adaptive Training Sample Selection

When training an object detector, we first need to define positive and negative samples for classification, and then use positive samples for regression. According to the previous analysis, the former one is crucial and the anchor-free detector FCOS improves this step. It introduces a new way to define positives and negatives, which achieves better performance than the traditional IoU-based strategy. Inspired by this, we **delve** into the most basic issue in object detection: how to define positive and negative training samples, and propose an Adaptive Training Sample Selection (ATSS). Compared with these traditional strategies, our method almost has no hyperparameters and is robust to different settings.

在训练物体检测器时，我们首先需要定义正负样本进行分类，然后用正样本进行回归。根据前面的分析，前者至关重要，无锚检测器FCOS改进了这一步骤。它引入了一种新的方式来定义阳性和阴性，比传统的基于IoU的策略取得了更好的性能。受此启发，我们深入研究了物体检测中最基本的问题：如何定义正负训练样本，并提出了自适应训练样本选择（ATSS）。与这些传统策略相比，我们的方法几乎没有超参数，并且对不同的设置具有鲁棒性。

4.1. Description

Previous sample selection strategies have some sensitive hyperparameters, such as IoU thresholds in anchor-based detectors and scale ranges in anchor-free detectors. After these hyperparameters are set, all ground-truth boxes must select their positive samples based on the fixed rules, which are suitable for most objects, but some outer objects will be neglected. Thus, different settings of these hyperparameters will have very different results. To this end, we propose the ATSS method that automatically divides positive and negative samples according to statistical characteristics of object almost without any hyperparameter. Algorithm 1 describes how the proposed method works for an input image.

For each ground-truth box g on the image, we first find out its candidate positive samples. As described in Line 3 to 6, on each pyramid level, we select k anchor boxes whose center are closest to the center of g based on L2 distance. Supposing there are L feature pyramid levels, the ground-truth box g will have k × L candidate positive samples.

After that, we compute the IoU between these candidates and the ground-truth g as Dg in Line 7, whose mean and standard deviation are computed as mg and vg in Line 8 and Line 9. With these statistics, the IoU threshold for this ground-truth g is obtained as tg = mg+vg in Line 10.

Finally, we select these candidates whose IoU are greater than or equal to the threshold tg as final positive samples in Line 11 to 15. Notably, we also limit the positive samples' center to the ground-truth box as shown in Line 12. Besides, if an anchor box is assigned to multiple ground-truth boxes, the one with the highest IoU will be selected. The rest are negative samples. Some motivations behind our method are explained as follows. Selecting candidates based on the center distance between anchor box and object. For RetinaNet, the IoU is larger when the center of anchor box is closer to the center of object. For FCOS, the closer anchor point to the center of object will produce higher-quality detections. **Thus, the closer anchor to the center of object is the better candidate.**

Algorithm 1 Adaptive Training Sample Selection (ATSS)

Input:

$\mathcal{G}$ is a set of ground-truth boxes on the image
$\mathcal{L}$ is the number of feature pyramid levels
$\mathcal{A}_i$ is a set of anchor boxes from the i_{th} pyramid levels
$\mathcal{A}$ is a set of all anchor boxes
k is a quite robust hyperparameter with a default value of 9

Output:

$\mathcal{P}$ is a set of positive samples
$\mathcal{N}$ is a set of negative samples

for each ground-truth $g \in \mathcal{G}$ **do**
  build an empty set for candidate positive samples of the ground-truth g: $\mathcal{C}_g \leftarrow \varnothing$;
  for each level $i \in [1, \mathcal{L}]$ **do**
    $\mathcal{S}_i \leftarrow$ select k anchors from A_i whose center are closest to the center of ground-truth g based on L2 distance;
    $\mathcal{C}_g = \mathcal{C}_g \cup \mathcal{S}_i$;
  end for
  compute IoU between $\mathcal{C}_g$ and g: $\mathcal{D}_g = IoU(\mathcal{C}_g, g)$;
  compute mean of $\mathcal{D}_g$: $m_g = Mean(\mathcal{D}_g)$;
  compute standard deviation of $\mathcal{D}_g$: $v_g = Std(\mathcal{D}_g)$;
  compute IoU threshold for ground-truth g: $t_g = m_g + v_g$;
  for each candidate $c \in \mathcal{C}_g$ **do**
    if $IoU(c, g) \geq t_g$ and center of c in g **then**
      $\mathcal{P} = \mathcal{P} \cup c$;
    end if
  end for
end for
$\mathcal{N} = \mathcal{A} - \mathcal{P}$;
return $\mathcal{P}, \mathcal{N}$;

Using the sum of mean and standard deviation as the IoU threshold.

The IoU mean mg of an object is **a measure of the suitability of the preset anchors for this object**. A high mg as shown in Figure 3(a) indicates it has high-quality candidates and the IoU threshold is supposed to be high. A low mg as shown in Figure 3(b) indicates that most of its candidates are low-quality and the IoU threshold should be low.

Besides, the IoU standard deviation vg of **an object is a measure of which layers are suitable to detect this object.** A high vg as shown in Figure 3(a) means there is a pyramid level specifically suitable for this object, adding vg to mg obtains a high threshold to select positives only from that level. A low vg as shown in Figure 3(b) means that there are several pyramid levels suitable for this object, adding vg to mg obtains a low threshold to select appropriate positives from these levels. Using the sum of mean mg and standard deviation vg as the IoU threshold tg can adaptively select enough positives for each object from appropriate pyramid levels in accordance of statistical characteristics of object.

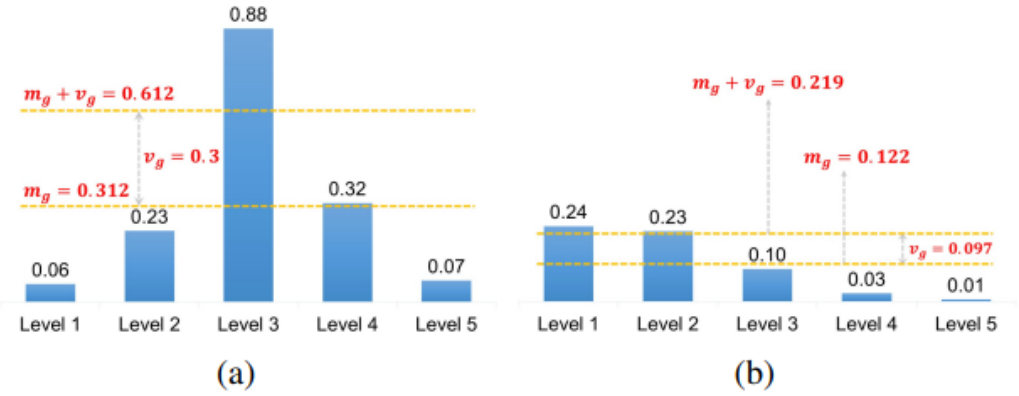


Figure 3: Illustration of ATSS. Each level has one candidate with its IoU. (a) A ground-truth with a high m_g and a high v_g. (b) A ground-truth with a low m_g and a low v_g.

Limiting the positive samples' center to object.

The anchor with a center outside object is a poor candidate and will be predicted by the features outside the object, which is not conducive to training and should be excluded Maintaining fairness between different objects. According to the statistical theory4 , about 16% of samples are in the confidence interval [mg + vg, 1] in theory. Although the IoU of candidates is not a standard normal distribution, **the statistical results show that each object has about 0.2 kL positive samples, which is invariant to its scale, aspect ratio and location. In contrast, strategies of RetinaNet and FCOS tend to have much more positive samples for larger objects, leading to unfairness between different objects.**

中心在物体之外的锚是一个很差的候选者，会被物体之外的特征所预测，这不利于训练，应该被排除在外 保持不同物体之间的公平性。根据统计学理论4，理论上约有16%的样本处于置信区间[mg + vg, 1]内。虽然候选人的IoU不是一个标准的正态分布，**但统计结果表明，每个对象都有大约0.2 kL的正样本，这对其比例、长宽比和位置是不变的。相比之下，RetinaNet和FCOS的策略往往对较大的物体有更多的正样本，导致不同物体之间的不公平。**

Keeping almost hyperparameter-free.

Our method only has one hyperparameter k. Subsequent experiments prove that it is quite insensitive to the variations of k and the proposed ATSS can be considered almost hyperparameter-free.

4.2. Verification

Anchor-based RetinaNet.

To verify the effectiveness of our adaptive training sample selection for anchor-based detectors, we use it to replace the traditional strategy in the improved RetinaNet (#A=1). As shown in Table 3, it consistently boosts the performance by 2.3% on AP, 2.4% on AP50, 2.9% for AP75, 2.9% for APS, 2.1% for APM and 2.7% for APL. These improvements are mainly due to the adaptive selection of positive samples for each ground-truth based on its statistical characteristics. Since

our method only redefines positive and negative samples without incurring any additional overhead, these improvements can be considered cost-free.

Anchor-free FCOS.

The proposed method can also be applied to the anchor-free FCOS in two different versions: the lite and full version. For the lite version, we apply some ideas of the proposed ATSS to FCOS, i.e., replacing its way to select candidate positives with the way in our method. FCOS considers anchor points in the object box as candidates, which results in plenty of low-quality positives. In contrast, our method selects top k = 9 candidates per pyramid level for each ground-truth. The lite version of our method has been merged to the official code of FCOS as the center sampling, which improves FCOS from 37.8% to38.6% on AP as listed in Table 3. However, the hyperparameters of scale ranges still exist in the lite version.

Table 3: Verification of the proposed method (%) on the MS COCO `minival` set. ATSS and center sampling are the full version and the lite version of our proposed method.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RetinaNet (#A=1)	37.0	55.1	39.9	21.4	41.2	48.6
RetinaNet (#A=1) + ATSS	39.3	57.5	42.8	24.3	43.3	51.3
FCOS	37.8	55.6	40.7	22.1	41.8	48.8
FCOS + Center sampling	38.6	57.4	41.4	22.3	42.5	49.8
FCOS + ATSS	39.2	57.3	42.4	22.7	43.1	51.5

For the full version, we let the anchor point in FCOS become the anchor box with 8S scale to define positive and negative samples, then still regress these positive samples to objects from the anchor point like FCOS. As shown in Table 3, it significantly increases the performance by 1.4% for AP, by 1.7% for AP50, by 1.7% for AP75, by 0.6% for APS, by 1.3% for APM and by 2.7% for APL. Notably, these two versions have the same candidates selected in the spatial dimension, but different ways to select final positives from candidates along the scale dimension. As listed in the last two rows of Table 3, the full version (ATSS) outperforms the lite version (center sampling) across different metrics by a large margin. These results indicate that the adaptive way in our method is better than the fixed way in FCOS to select positives from candidates along the scale dimension.

4.3. Analysis

Training an object detector with the proposed adaptive training sample selection only involves one hyperparameter k and one related setting of anchor boxes. This subsection analyzes them one after another.

Hyperparameter k.

We conduct several experiments to study the robustness of the hyperparameter k, which is used to select the candidate positive samples from each pyramid level. As shown in Table 4, different values of k in [3, 5, 7, 9, 11, 13, 15, 17, 19] are used to train the detector. We observe that the proposed method is quite insensitive to the variations of k from 7 to 17. Too large k (e.g., 19) will result in too many low-quality candidates that slightly decreases the performance. Too small k (e.g., 3) causes a noticeable drop in accuracy, because too few candidate positive samples will

cause statistical instability. Overall, the only hyperparameter k is quite robust and the proposed ATSS can be nearly regarded as hyperparameter-free.

Table 4: Analysis of different values of hyperparameter k on the MS COCO `minival` set.

k	3	5	7	9	11	13	15	17	19
AP (%)	38.0	38.8	39.1	39.3	39.1	39.0	39.1	39.2	38.9

Anchor Size.

The introduced method resorts to the anchor boxes to define positives and we also study the effect of the anchor size. In the previous experiments, one square anchor with 8S (S indicates the total stride size of the pyramid level) is tiled per location. As shown in Table 5, we conduct some experiments with different scales of the square anchor in [5, 6, 7, 8, 9] and the performances are quite stable. Besides, several experiments with different aspect ratios of the 8S anchor box are performed as shown in Table 6. The performances are also insensitive to this variation. These results indicate that the proposed method is robust to different anchor settings.

Table 5: Analysis (%) of different anchor scales with fixed aspect ratio $1:1$ on the MS COCO `minival` set.

Scale	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
5	39.0	57.9	41.9	23.2	42.8	50.5
6	39.2	57.6	42.5	23.5	42.8	51.1
7	39.3	57.6	42.4	22.9	43.2	51.3
8	39.3	57.5	42.8	24.3	43.3	51.3
9	38.9	56.5	42.0	22.9	42.4	50.3

4.4. Comparison

We compare our final models on the MS COCO test-dev subset in Table 8 with other state-of-the-art object detectors. Following previous works [33, 56], the multiscale training strategy is adopted for these experiments, i.e., randomly selecting a scale between 640 to 800 to resize the shorter side of images during training. Besides, we double the total number of iterations to 180K and the learning rate reduction points to 120K and 160K correspondingly. Other settings are consistent with those mentioned before.

As shown in Table 8, our method with ResNet-101 achieves 43.6% AP without any bells and whistles, which is better than all the methods with the same backbone including Cascade R-CNN [5] (42.8% AP), C-Mask RCNN [7] (42.0% AP), RetinaNet [33] (39.1% AP) and RefineDet [66] (36.4% AP). We can further improve the AP accuracy of the proposed method to 45.1% and 45.6% by using larger backbone networks ResNeXt-32x8d-101 and ResNeXt-64x4d-101 [63], respectively. The 45.6% AP result surpasses all the anchor-free and anchor-based de□tectors except only 0.1% lower than SNIP [54] (45.7% AP), which introduces the improved multi-scale trainingand testing strategy. Since our method is about the definition of positive and negative samples, it is

compatible and complementary to most of current technologies. We further use the Deformable Convolutional Networks (DCN) [10] to the ResNet and ResNeXt backbones as well as the last layer of detector towers. DCN consistently improves the AP performances to 46.3% for ResNet-101, 47.7% for ResNeXt-32x8d-101 and 47.7% for ResNeXt-64x4d-101, respectively. The best result 47.7% is achieved with single-model and single-scale testing, outperforming all the previous detectors by a large margin. Finally, with the multi-scale testing strategy, our best model achieves 50.7% AP.

Table 8: Detection results (%) on MS COCO `test-dev` set. Bold fonts indicate the best performance.

Method	Data	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
anchor-based two-stage:								
MLKP [58]	trainval35	ResNet-101	28.6	52.4	31.6	10.8	33.4	45.1
R-FCN [9]	trainval	ResNet-101	29.9	51.9	-	10.8	32.8	45.0
CoupleNet [74]	trainval	ResNet-101	34.4	54.8	37.2	13.4	38.1	50.8
TDM [53]	trainval	Inception-ResNet-v2-TDM	36.8	57.7	39.2	16.2	39.8	52.1
Hu et al. [18]	trainval35k	ResNet-101	39.0	58.6	42.9	-	-	-
DeepRegionlets [64]	trainval35k	ResNet-101	39.3	59.8	-	21.7	43.7	50.9
FitnessNMS [57]	trainval	DeNet-101	39.5	58.0	42.6	18.9	43.5	54.1
Gu et al. [15]	trainval35k	ResNet-101	39.9	63.1	43.1	22.2	43.4	51.6
DetNet [31]	trainval35k	DetNet-59	40.3	62.1	43.8	23.6	42.6	50.0
Soft-NMS [3]	trainval	ResNet-101	40.8	62.4	44.9	23.0	43.4	53.2
SOD-MTGAN [1]	trainval35k	ResNet-101	41.4	63.2	45.4	24.7	44.2	52.6
G-RMI [19]	trainval35k	Ensemble of Five Models	41.6	61.9	45.4	23.9	43.5	54.9
C-Mask RCNN [7]	trainval35k	ResNet-101	42.0	62.9	46.4	23.4	44.7	53.8
Cascade R-CNN [5]	trainval35k	ResNet-101	42.8	62.1	46.3	23.7	45.5	55.2
Revisiting RCNN [8]	trainval35k	ResNet-101+ResNet-152	43.1	66.1	47.3	25.8	45.9	55.3
SNIP [54]	trainval35k	DPN-98	45.7	67.3	51.1	29.3	48.8	57.1
anchor-based one-stage:								
YOLOv2 [46]	trainval35k	DarkNet-19	21.6	44.0	19.2	5.0	22.4	35.5
SSD512* [36]	trainval35k	VGG-16	28.8	48.5	30.3	10.9	31.8	43.5
STDN513 [69]	trainval	DenseNet-169	31.8	51.0	33.6	14.4	36.1	43.4
DES512 [68]	trainval35k	VGG-16	32.8	53.2	34.5	13.9	36.2	47.5
DSSD513 [12]	trainval35k	ResNet-101	33.2	53.3	35.2	13.0	35.4	51.1
RFB512-E [35]	trainval35k	VGG-16	34.4	55.7	36.4	17.6	37.0	47.6
PFPNet-R512 [21]	trainval35k	VGG-16	35.2	57.6	37.9	18.7	38.6	45.9
RefineDet512 [66]	trainval35k	ResNet-101	36.4	57.5	39.5	16.6	39.9	51.4
RetinaNet [33]	trainval35k	ResNet-101	39.1	59.1	42.3	21.8	42.7	50.2
anchor-free keypoint-based:								
ExtremeNet [71]	trainval35k	Hourglass-104	40.2	55.5	43.2	20.4	43.2	53.1
CornerNet [26]	trainval35k	Hourglass-104	40.5	56.5	43.1	19.4	42.7	53.9
CenterNet-HG [70]	trainval35k	Hourglass-104	42.1	61.1	45.9	24.1	45.5	52.8
Grid R-CNN [39]	trainval35k	ResNeXt-101	43.2	63.0	46.6	25.1	46.5	55.2
CornerNet-Lite [27]	trainval35k	Hourglass-54	43.2	-	-	24.4	44.6	57.3
CenterNet [11]	trainval35k	Hourglass-104	44.9	62.4	48.1	25.6	47.4	57.4
RepPoints [65]	trainval35k	ResNet-101-DCN	45.0	66.1	49.0	26.6	48.6	57.5
anchor-free center-based:								
GA-RPN [59]	trainval35k	ResNet-50	39.8	59.2	43.5	21.8	42.6	50.7
FoveaBox [23]	trainval35k	ResNeXt-101	42.1	61.9	45.2	24.9	46.8	55.6
FSAF [72]	trainval35k	ResNeXt-64x4d-101	42.9	63.8	46.3	26.6	46.2	52.7
FCOS [56]	trainval35k	ResNeXt-64x4d-101	43.2	62.8	46.6	26.5	46.2	53.3
Ours:								
ATSS	trainval35k	ResNet-101	43.6	62.1	47.4	26.1	47.0	53.6
ATSS	trainval35k	ResNeXt-32x8d-101	45.1	63.9	49.1	27.9	48.2	54.6
ATSS	trainval35k	ResNeXt-64x4d-101	45.6	64.6	49.7	28.5	48.9	55.6
ATSS	trainval35k	ResNet-101-DCN	46.3	64.7	50.4	27.7	49.8	58.4
ATSS	trainval35k	ResNeXt-32x8d-101-DCN	47.7	66.6	52.1	29.3	50.8	59.7
ATSS	trainval35k	ResNeXt-64x4d-101-DCN	47.7	66.5	51.9	29.7	50.8	59.4
ATSS (Multi-scale testing)	trainval35k	ResNeXt-32x8d-101-DCN	50.6	68.6	56.1	**33.6**	**52.9**	62.2
ATSS (Multi-scale testing)	trainval35k	ResNeXt-64x4d-101-DCN	**50.7**	**68.9**	**56.3**	33.2	**52.9**	**62.4**

4.5. Discussion

Previous experiments are based on RetinaNet with only one anchor per location. There is still a difference between anchor-based and anchor-free detectors that is not explored: the number of anchors tiled per location. Actually, the orig□inal RetinaNet tiles 9 anchors (3 scales × 3 aspect ratios) per location (marked as RetinaNet (#A=9)) that achieves 36.3% AP as listed in the first row of Table 7. In addition, those universal improvements in Table 1 can also be used to RetinaNet (#A=9), boosting the AP performance from 36.3% to 38.4%. Without using the proposed ATSS, the improved RetinaNet (#A=9) has better performance than RetinaNet (#A=1), i.e., 38.4% in Table 7 vs. 37.0% in Table 1. These results indicate that under the traditional IoU-based sample selection strategy, tiling more anchor boxer per location is effective.

Table 7: Results (%) with different multiple anchors per location on the MS COCO `minival` set.

Method	#sc	#ar	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RetinaNet (#A=9)	3	3	36.3	55.2	38.8	19.8	39.8	48.8
+Imprs.	3	3	38.4	56.2	41.6	22.2	42.4	50.1
+Imprs.+ATSS	3	3	39.2	57.6	42.7	23.8	42.8	50.9
+Imprs.+ATSS	3	1	39.3	57.7	42.6	23.8	43.5	51.2
+Imprs.+ATSS	1	3	39.2	57.1	42.5	23.2	43.1	50.3
+Imprs.+ATSS	1	1	39.3	57.5	42.8	24.3	43.3	51.3

Table 1: Analysis of implementation inconsistencies between RetinaNet and FCOS on MS COCO `minival` set. "#A=1" means there is one square anchor box per location.

Inconsistency	FCOS	RetinaNet (#A=1)					
GroupNorm	✓		✓	✓	✓	✓	✓
GIoU Loss	✓			✓	✓	✓	✓
In GT Box	✓				✓	✓	✓
Centerness	✓					✓	✓
Scalar	✓						✓
AP (%)	37.8	32.5	33.4	34.9	35.3	36.8	37.0

However, after using our proposed method, the opposite conclusion will be drawn. To be specific, the proposed ATSS also improves RetinaNet (#A=9) by 0.8% on AP, 1.4% on AP50 and 1.1% on AP75, achieving similar performances to RetinaNet (#A=1) as listed in the third and sixth rows of Table 7. Besides, when we change the number of anchor scales or aspect ratios from 3 to 1, the results are almost unchanged as listed in the fourth and fifth rows of Table 7. In other words, as long as the positive samples are selected appropriately, no matter how many anchors are tiled at each location, the results are the same. We argue that tiling multiple anchors per location is a useless operation under our proposed method and it needs further study to discover its right role.

5. Conclusion

In this work, we point out that the essential difference be□tween one-stage anchor-based and center-based anchor-free detectors is actually the definition of positive and negative training samples. It indicates that how to select positive and negative samples during object detection training is critical. Inspired by that, we delve into this basic issue and propose the adaptive training sample selection, which automatically divides positive and negative training samples according to statistical characteristics of object, hence bridging the gap between anchor-based and anchor-free detectors. We also discuss the necessity of tiling multiple anchors per location and show that it may not be a so useful operation under current situations. Extensive

experiments on the challenging benchmarks MS COCO illustrate that the proposed method can achieve state-of-the-art performances without introducing any additional overhead.