激光与光电子学进展 先进成像

基于语义采样和检测框优化的目标检测算法

李昱[1,2]，盖绍彦[1,2,3]，达飞鹏[1,2,3*]，洪儒[1,2]

[1]东南大学自动化学院，江苏 南京 210096；

[2]东南大学复杂工程系统测量与控制教育部重点实验室，江苏 南京 210096；

[3]东南大学深圳研究院，广东 深圳 518063

摘要 样本采样和检测框优化是目标检测任务中的两项重要技术。为了解决正负样本分配不合理的问题，获取更优的图像分类特征和检测框，提出一个精确且高效的单阶无锚框目标检测算法，算法由基于语义的定位、自适应特征增强和高效的检测框优化3个模块组成。首先，定位模块提出基于语义的样本采样方法，根据目标的语义特征区分前/背景区域，合理选择正样本和负样本，优先选择语义信息量较大的前景区域作为正样本；其次，特征增强模块利用目标语义概率图和检测框偏移逐像素调整图像分类特征，增大前景特征所占比重，根据目标大小自适应调整特征编码范围；最后，采用并联的方式优化检测框，对优化前后的检测框计算分类损失，几乎无成本地提升了定位性能，保证了特征对齐性和一致性。在MS COCO数据集下，提出的目标检测算法取得了平均精度为42.8% 的检测精度，单张图像的检测时间达到78 ms，实现了检测精度与速度的平衡。

关键词 机器视觉；目标检测；正负样本采样；检测框优化；特征增强

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP202259.1815015

Object Detection Based on Semantic Sampling and Localization Refinement

Li Yu[1,2], Gai Shaoyan[1,2,3], Da Feipeng[1,2,3*], Hong Ru[1,2]

[1]*School of Automation, Southeast University, Nanjing* 210096, *Jiangsu, China*;

[2]*Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Southeast University, Nanjing* 210096, *Jiangsu, China*;

[3]*Shenzhen Research Institute, Southeast University, Shenzhen* 518063, *Guangdong, China*

Abstract The two important techniques for object detection are training samplers and localization refinement. To solve the problem of unreasonable distribution of positive and negative samples, and get better image classification features and localizations, this study presented an accurate and effective single step anchor-free algorithm for object detection. The algorithm consists of three modules: semantic based positioning, adaptive feature enhancement, and efficient localization refinement. Firstly, the positioning module proposes a semantic based sampling method, which distinguishes the front/background regions according to the semantic characteristics of the object, reasonably selects positive samples and negative samples, and preferentially selects the foreground region with large amount of semantic information as the positive samples. Secondly, the feature enhancement module uses the target semantic probability map and detection frame offset to adjust the image classification features pixel by pixel, increases the proportion of foreground features, and adaptively adjusts the feature coding range according to the object size. Finally, the localizations are optimized in parallel, and the classification loss is calculated for the localizations before and after optimization, which improves the positioning performance almost without cost, and ensures the feature alignment and consistency. In the MS COCO dataset, the proposed algorithm achieves 42.8% in average precision, the detection time of a single image reaches 78 ms, realizing the balance between detection accuracy and speed.

Key words machine vision; object detection; positive and negative training sampler; localization refinement; feature enhancement

收稿日期：2021-07-23；**修回日期**：2021-08-25；**录用日期**：2021-08-31

基金项目：国家自然科学基金(51475092)、江苏省自然科学基金(BK20181269)、江苏省前沿引领技术基础研究专项(BK20192004C)、深圳市科技创新委员会(JCYJ20180306174455080)

通信作者：*qxxymm@163.com

1 引 言

目标检测是许多计算机视觉任务的重要基础，在自动驾驶[1]、行为识别[2]和遥感图像检测[3]等方面都有广泛的应用价值。随着社交媒体平台的激增，大量的视觉信息需要被处理，智能系统对目标检测的性能要求越来越高，尤其是在复杂场景中，需要同时处理多个感兴趣目标，对目标进行快速且精确的检测就显得尤为重要。

基于卷积神经网络的目标检测算法有多种类型，根据是否预设锚框可分为有锚框和无锚框两种。从CornerNet算法[4]开始，无锚框目标检测算法层出不穷，如FSAF[5]、FCOS[6]、ExtremeNet[7]等算法。无锚框检测模型消除了与锚框相关的计算和超参数，得到了学术界越来越多的关注和认可。无锚框目标检测器通过两种不同的方式寻找目标：一种是基于关键点的方式，即先确定几个预定义的关键点，如角点[4]、极值点[7]、中心点[8]等，再根据人工制定的规则将这些关键点组合起来生成检测框；另一种是基于中心的方式，用目标的中心点或中心区域来定义正样本，然后预测正样本到目标边界的距离，如FCOS[7]、FoveaBox[9]等算法。然而，基于关键点的目标检测方法受到人工聚类规则和后处理方法的限制。在基于中心点的目标检测算法中，初始正样本点至关重要，需要训练样本采样器来合理选择正、负样本。FCOS算法[7]根据中心度降低远离目标中心的样本点分数，优先选择靠近目标中心的样本；Foveabox算法[9]引入两个缩放因子将靠近目标中心的区域当作正样本，远离中心的区域当作负样本，忽略中间区域；ATSS算法[10]优先选择靠近目标中心的k个样本点作为正样本，再根据目标的统计特征自动选择正样本和负样本。利用这些训练样本采样方法一定程度上提升了检测性能，但优先选择靠近目标中心的样本点为正样本是不合理的，因为各类目标有不同的形状特征，有些目标的中心会落在背景或者语义信息很少的前景上，这些靠近目标中心的样本点包含很少的前景特征，将其作为正样本，反而不利于模型训练。

为了达到更高的检测精度，许多基于检测框优化的目标检测算法被提出。Cascade RCNN[11]和HTC[12]等级联式优化算法通过对检测框的多次分类和回归大大提高了检测精度，检测速度往往较慢。单阶目标检测算法RefineDet[13]也运用了类似的检测框优化思想来提高检测精度。与使用候选框池化(RoI Pooling)或候选框对齐(RoI Align)进行特征对齐的双阶目标检测算法不同，基于级联式候选框优化的单阶检测器还不能很好地保证特征对齐。RepPoints算法[14]使用可变形卷积进行特征对齐，但通过学习检测框偏移进行特征对齐的方法过于隐式，不能保证特征是真正对齐的。R^3Det算法[15]通过计算找到检测框优化前后对应的特征区域，再重建特征图实现特征对齐，增加了额外的计算量。Kong等[16]指出单阶目标检测算法的双分支结构还导致了特征的不一致问题，分类分支预测原始检测框类别概率，而定位分支将原始检测框转化为更接近目标真值的检测框，当原始检测框与优化后的检测框之间存在较大差异时，特征不对齐和特征不一致问题变得更加突出。

针对以上问题，本文提出一种基于语义采样和检测框优化的单阶无锚框目标检测算法，主要由3个模块组成：基于语义的定位模块、自适应特征增强模块和检测框优化模块。首先，基于语义的定位模块能够更合理地分配正、负样本，优先选择富含前景信息的样本点作为目标的初始表示，同时自顶向下自动学习表示目标的一系列关键点，不需要人工聚类步骤；其次，为了增强分类特征，使其更加关注目标实例的区域，特征增强模块利用语义信息逐像素调整特征比重，自适应调整特征编码范围，使大目标拥有大范围的编码特征，而小目标对应较小的特征范围；最后，检测框优化模块采用并联的方式优化检测框，相比级联式优化方法，提升了检测速度，同时保证了检测框优化过程中的特征对齐性和一致性。在MS COCO[17]和PASCAL VOC[18]数据集上验证了本文提出的目标检测算法及各个模块的有效性，且算法实现了检测精度与速度的平衡。

2 基本原理和方法

本节首先回顾先进的无锚框目标检测算法RepPoints[14]，它将目标表示为自上而下自动学习到的一系列关键点(点集)。然后介绍本文提出的基于语义采样和检测框优化的单阶无锚框目标检测算法总体架构。最后详细阐述基于语义的定位模块、自适应特征增强模块和检测框优化模块的实现细节。

2.1 RepPoints

RepPoints[14]是基于中心点的无锚框目标检测算法，该算法用点集代替传统的矩形框表示目标，自上而下自动学习表示目标的关键点。如图1(a)所示，RepPoints定位分支以级联的方式优化各个关键点的位置以获取精细的目标定位，样本点P_0为初始目标表示，从该样本点开始回归点集R_1作为初始候选点集，再利用可变形卷积进行第二阶段候选点集优化，得到最终优化点集R_2；在分类分支中，RepPoints根据定位分支第一阶段的偏移量自适应调整分类特征，只对原始点集R_1计算分类损失L_c，公式如下：

$$L_c=\frac{1}{N_{cls}}\sum_j L_{cls}\left(c_j,c_j^*\right),\tag{1}$$

式中：L_{cls}为分类损失函数；N_{cls}表示批量大小；c_j和c_j^*分别表示第j个点集为目标的预测概率和对应的分类标签。

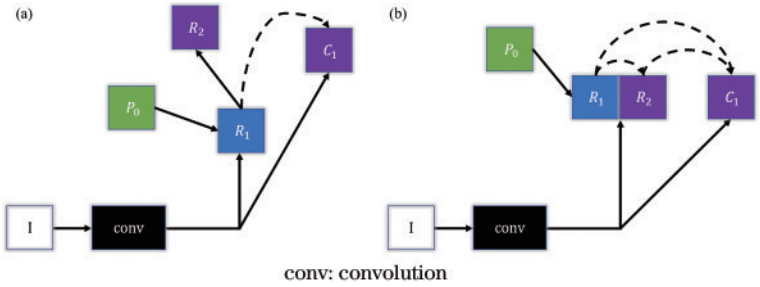

图 1　目标检测器架构对比图。(a) RepPoints[14]架构图;(b) 提出目标检测架构图

Fig. 1　Comparison of object detection architectures. (a) Architecture of RepPoints[14]; (b) architecture of proposed detector

相比传统的矩形框表示目标的方法,RepPoints[14]利用关键点集合表示目标能够获取更精细的检测框位置,然而其检测框定位严重依赖回归,这意味着目标的初始表示和优化极其重要,而选择距离目标真值中心最近的k个样本点作为正样本,初始样本点P_0很可能落在背景区域或者目标语义信息较少的前景区域,制约了定位的准确性。其次,RepPoints级联式的检测框优化在一定程度上限制了算法的检测效率,当原始点集和优化后的点集之间存在较大差异时,会导致特征不对齐和不一致问题。本文提出的检测算法架构图如图 1(b)所示,在定位时,先根据语义信息引导初始样本点P_0的选择,以并联的方式一次性预测点集R_1和R_2,初始点集R_1的空间偏移用来指导优化点集R_2的生成;分类分支利用定位分支产生的目标语义特征和空间偏移自适应增强分类特征,同时利用点集R_1和R_2共同监督图像分类。实验证明本文目标检测算法可以有效地提高检测性能。

2.2　整体网络架构

图 2 展示了本文提出的目标检测网络整体架构图,检测框架由基于语义的定位模块和自适应特征增强模组成,高效的检测框优化方法见 2.5 节。给定一幅图像,先获得一层图像特征图F_I,它是图像经FPN[19]骨干网络输出的一层特征映射。骨干网络连接两个非共享子网络,分别用于目标定位和图像分类,图 3 展示了定位子网络和分类子网络结构图。定位子网络首先应用 3 层 256 通道 3×3 卷积层(3×3 conv, 256)得到定位特征图F_R,然后经过基于语义的定位模块,如图 2 虚线框所示,该模块包含两个分支,分别得到位置可能性特征图F_L和最终目标定位特征图F'_R。F_L表明各个位置包含目标语义信息量大小。点集预测分支自动学习表示目标的一系列关键点R,包含R_1、R_2两部分内容。分类子网络同样应用 3 层 3×3 conv,256 获得初始分类特征图F_C,再通过特征增强模块得到最终的分类特征图F'_C。如图 2 实线框所示,特征增强模块利用定位分支中产生的二值概率图O_r和

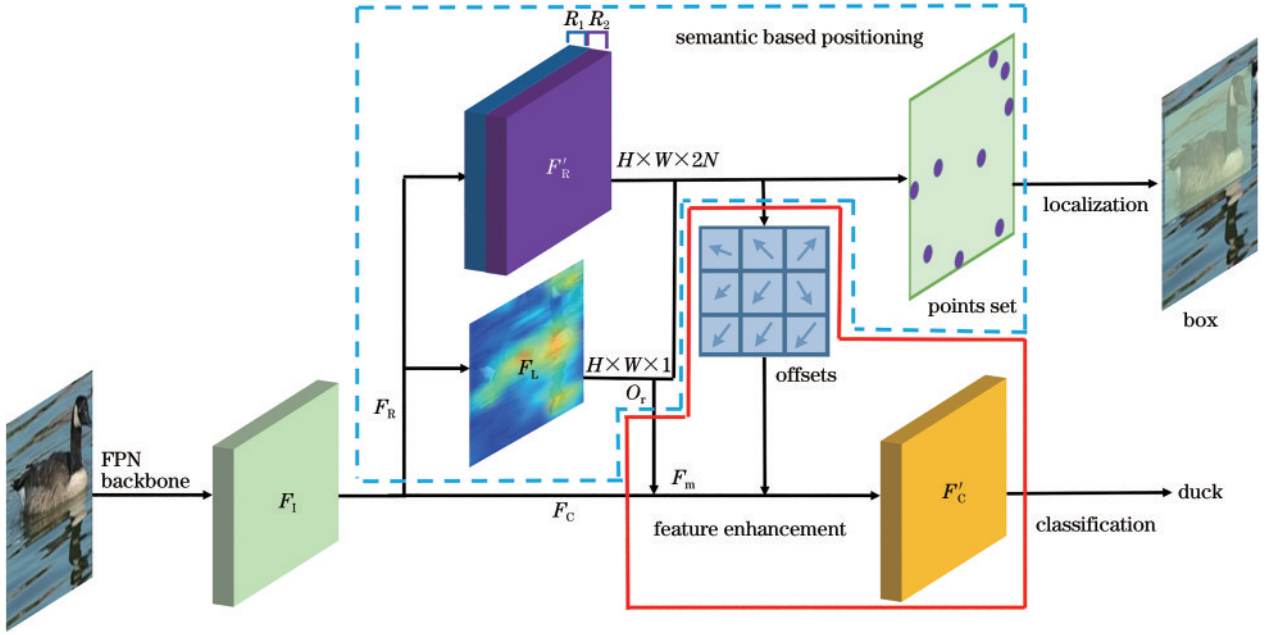

图 2　提出的整体网络架构图。虚线框:基于语义的定位模块;实线框:特征增强模块

Fig. 2　Overall architecture of proposed object detector. Dashed box: semantic based positioning module; dashed box: feature enhancement module

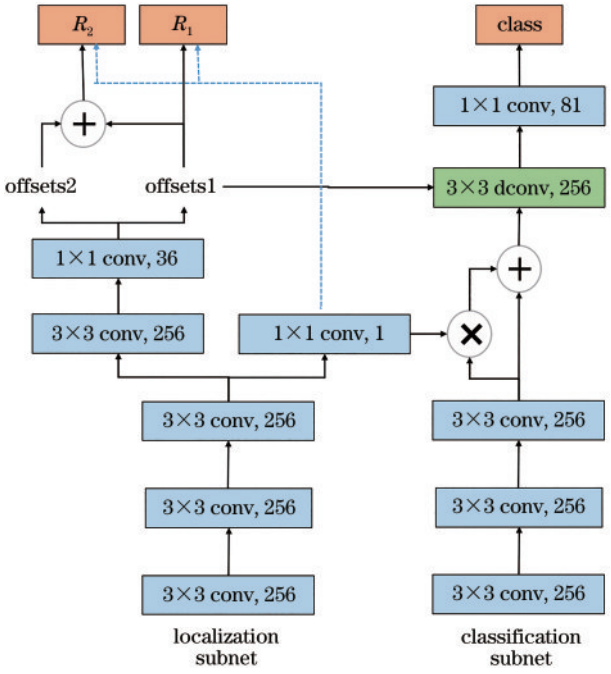

图 3　定位子网络和分类子网络结构图

Fig. 3　Architecture of localization subnet and classification subnet

点集偏移量(offsets)来增强分类特征图 F_C。最后，利用并联式检测框优化方法，提高了检测速度的同时解决了目标检测算法中存在的不对齐和不一致问题(详见 2.5 节)。

与 RepPoints[14] 相比，提出的检测框架增加了位置信息预测分支，利用目标语义信息获得各个位置存在目标的可能性特征图 F_L，表示每个位置包含目标语义信息量大小，并依此选择正、负样本点，优先选择语义信息丰富的正样本点作为目标的初始表示 P_0，而不是单纯依据样本点到目标真值中心的距离选择正样本。特征增强模块不仅利用定位分支产生的 offsets 来捕获不同视觉范围内的内容，还利用根据目标前景信息权重 O_r 逐像素调整特征比重，使分类特征更加关注包含目标实例的区域。最后，提出的检测框优化方法，在训练时巧妙地优化了检测框，比 RepPoints 级联式的检测框优化方式运行效率更高，同时保证了检测框优化过程中的特征对齐性和一致性。

2.3　基于语义的定位模块

如图 2 中的虚线框所示，基于语义的定位模块由位置信息预测分支和点集预测分支组成，分别预测目标语义信息量的特征图 F_L 和定位特征图 F'_R。该模块提出基于语义的训练样本采样器来分配正、负样本，利用语义信息获取初始的目标表示。样本采样分两个步骤实现：1)获取语义信息量概率图 O_r；2)选取合适的样本点作为正样本。

首先，位置信息预测分支产生语义信息量概率图 O_r，其大小与输入定位特征图 F_R 大小相同。每个位置的值 $O_r(x,y)$ 表示像素点 (x,y) 包含语义信息的概率，$O_r(x,y)$ 值越高，则该像素包含的语义信息越丰富，选择该点作为训练正样本，可以获得更好的目标表示。语义概率图 O_r 通过连续两个步骤获得

$$F_L(x,y)=N_L(x,y,F_R), \tag{2}$$

$$O_r(x,y)=\sigma\left[F_L(x,y)\right], \tag{3}$$

式中：σ 表示 Sigmoid 激活函数；N_L 为 $F_R \to F_L$ 的转换函数。转换函数可以由几种不同的方案实现：1×1 conv、沿通道方向的最大池化(max-pooling)、沿通道方向组合最大池化和平均池化(max-pooling+avg-pooling)。根据经验，只使用一个转换层即可保持检测效率和精度之间的平衡。

其次，根据语义信息量概率图 O_r，选择合适的训练正样本。基于中心的样本采样器根据样本点与目标真值中心的距离选择正负样本，如 FCOS[6]、FoveaBox[9] 和 RepPoints[14]，这些算法均假设靠近目标中心的样本点是更好的选择。如图 4 所示，虚线框为目标真值，四角形表示目标真值中心，圆点和“*”号分别表示采样器选择到的正、负样本点。假设选择 k 个正样本点，图 4(a)基于中心的样本采样器优先选择靠近真值中心的 k 个点作为正样本，而许多正样本点落在语义信息量很小的位置，甚至落在背景区域，而包含更多语义信息量的样本点却被忽略。为此，本文提出基于语义的训练样本采样器，如图 4(b)所示，根据采样点包含的语义信息量，选择在目标真值框内语义信息量较大的 k 个点为正样本，即根据生成的目标概率图 O_r，选择满足以下条件的 k 个样本点作为训练正样

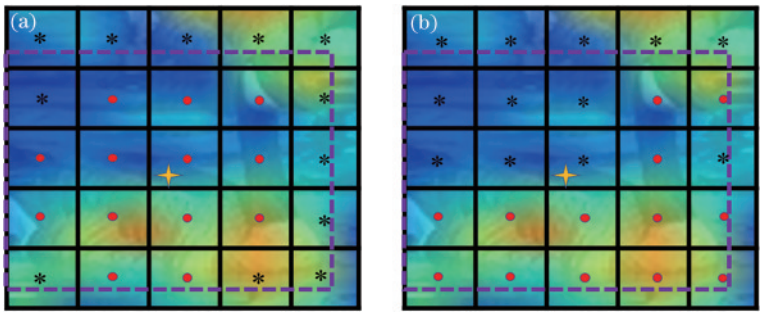

图 4　训练样本采样器。(a)基于中心的训练样本采样器；(b)基于语义的训练样本采样器

Fig. 4　Training samplers for object detection. (a) Center-based training sampler; (b) semantic-based training sampler

本：1）样本点位于目标真值框内；2）样本点的语义信息量高于预定义阈值ϵ_L；3）按语义信息量大小排序，取前k个点为正样本点。

点集预测分支将包含足够语义信息的特征点作为目标的初始表示，自顶向下自动学习表示目标的点集。RepPoints点集预测通道数为N，其中$N=2n$，n为点集中关键点数目，本文算法中，该分支的通道数是RepPoints的两倍，即$2N$，一半用于生成原始点集，另一半用于生成优化点集，这个过程将在2.5节中详细说明。

2.4 自适应特性增强模块

为了获取更好的分类特征，使其更加关注目标实例的区域，特征增强模块利用的语义信息逐像素调整特征比重，同时自适应调整特征编码范围，使得大目标拥有大范围的编码特征，小目标对应较小的特征范围。该模块由两部分组成：基于语义的特征增强和基于空间偏移的特征自适应，如图2中的实线框所示。

1）基于语义的特征增强。由于不同像素点包含不同的语义信息量，语义信息丰富的像素相比语义信息量较少的像素应在分类中起更大作用。基于语义的特征增强根据二值概率图权重，逐个像素增强原分类特征F_C。公式如下：

$$F_m = F_C \odot O_r + F_C, \tag{4}$$

式中：F_m是增强后的特征；$\odot$表示像素级点乘。

2）基于空间偏移的特征自适应。被检测目标大小不同，对应的视觉范围也应有所不同，大物体应该对大区域的内容进行编码，小目标应限定在较小的特征范围内，因此根据目标形状自适应变换特征，公式如下：

$$F'_C = f_{3\times3}[F_m, \Delta p_{1i}(x,y)_{i=1}^{n}], \tag{5}$$

式中：F'_C为自适应变换后的分类特征；$f_{3\times3}$为3×3的可变形卷积；F_m是基于语义增强后的特征；$\Delta p_{1i}(x,y)_{i=1}^{n}=(\Delta x_{1i},\Delta y_{1i})_{i=1}^{n}$表示$(x,y)$位置的核偏移量，即该位置相应的$n$个偏移量。如图3所示，先从点集预测分支预测一个偏移量offsets1，再作为可变形卷积的输入自适应调整分类特征。利用自适应特性增强模块，网络可以根据每个位置的语义信息量和视觉范围逐像素增强分类特征，使分类特征更加关注目标实例区域的特征。

2.5 高效的检测框优化

如图1(a)所示，目标检测器通常会像RepPoints[14]采用级联方式来优化检测框。然而，级联式检测框优化方法制约了整个检测器的运行速度，还会造成特征不对齐和不一致的问题。为了解决这些问题，本文提出一种高效的检测框优化方法，如图1(b)所示，一次性预测点集R_1和R_2，两者共同监督图像分类。如图2、图3所示，点集预测分支得到目标定位特征图F'_R，一半特征预测原始点集R_1相对于P_0的偏移offsets1，一半特征预测优化点集R_2相对于原始点集R_1的偏移offsets2，通过加法运算获得优化点集R_2。从初始正样本点$P_0=(x,y)$起始，定位过程通过两个连续的步骤实现：

$$R_1=\{p_{1i}=(x_{1i},y_{1i})\}_{i=1}^{n}=\{p_{1i}=(P_0+\Delta p_{1i})\}_{i=1}^{n}=\{(x+\Delta x_{1i},y+\Delta y_{1i})\}_{i=1}^{n}, \tag{6}$$

$$R_2=\{p_{2i}=(x_{2i},y_{2i})\}_{i=1}^{n}=\{p_{1i}=(p_{1i}+\Delta p_{2i})\}_{i=1}^{n}=\{p_{2i}=(x_{1i}+\Delta x_{2i},y_{1i}+\Delta y_{2i})\}_{i=1}^{n}, \tag{7}$$

式中：p_{1i}，p_{2i}分别表示点集R_1，R_2的第i个关键点；n为每个点集中关键点总数，默认为9。(x,y)、(x_{1i},y_{1i})、(x_{2i},y_{2i})分别表示P_0、p_{1i}、p_{2i}点的坐标；$\Delta p_{1i}=(\Delta x_{1i},\Delta y_{1i})$是$p_{1i}$相对于初始样本点$P_0$的偏移offsets1；$\Delta p_{2i}=(\Delta x_{2i},\Delta y_{2i})$是$p_{2i}$相对于$p_{1i}$的偏移offsets2。由于一次性预测不会改变原始点集和优化点集的样本点位置，保证了特征的绝对对齐。其次，在原有的分类损失中利用优化点集来共同监督图像分类，缓解了点集优化前后差异造成的不一致问题，分类损失函数为

$$L_c=\frac{1}{N_{cls}}\sum_j\left[L_{cls}(c_j,c_j^*)+\alpha L_{cls}(c_j,c_j^\dagger)\right], \tag{8}$$

式中：L_{cls}是分类损失；N_{cls}表示批量大小；c_j为优化点集被预测为目标的概率；c_j^*和$c_j^\dagger$分别表示原始点集和优化点集对应的目标真值；α是用于平衡两项的权重，取0.5。可以发现，将两者结合起来可以使训练过程更加稳定，并取得更好的检测结果。

在训练阶段，原始点集和优化点集都用来训练目标定位和分类。通过并联分块的方式，去掉了额外的级联子网络，提高了检测效率。在运行阶段，检测框定位和分类结果的生成过程与单阶检测器相同，无成本地提高了检测精度。

3 实验结果与分析

3.1 实验环境

1）数据集和评价指标。在最具挑战性的MS COCO[17]数据集上评估本文目标检测算法的性能，并在PASCAL VOC 2007[18]数据集上测试该算法的检测效果。在训练阶段，所有模型都在train2017上训练，它包含了80类目标的118 k图像，在5 k图像的val2017进行消融实验，在20 k图像的test-dev测试检测性能。检测精度采用COCO数据集评价指标AP(Average Precision)来评估实验结果。检测效率依据性能指标，如参数量(Param)、计算量(FLOPs)、帧率(FPS)和检测时间(Time)，这些都是影响能源消耗和网络效率的重要指标。

2）实验细节。所有模型都基于开源的mmdetection[20]代码库实现，除特别说明，消融实验均使用ResNet50-FPN[19]骨干网络。在图像处理方面，除了标准的水平图像翻转外，不使用其他数据增强，同时将所有图像不改变长宽比缩放到1333 pixel×800 pixel

大小。在2个图形处理器(GPU)上训练模型,每个批量2张图片,模型优化方法采用随机梯度下降(SGD),依据“1×”设置学习率,训练12轮(epoch),检测时间是在RTX2070 GPU上测量的。

3.2 实验结果

表1展示了多个目标检测算法在MS COCO[17]数据集的检测结果,所有模型均在单模型单尺度下训练得到。采用ResNet101-FPN[19]骨干网络时,本文提出的检测算法在COCO基准上AP达到42.8,单张图像的检测时间为78 ms。与基准网络RepPoints相比,检测精度提高4%,运行速度提高10%。与其他目标检测算法比,提出的目标检测算法取得了最快的检测速度,检测精度超过了大多目标检测算法。与最先进的检测器LSNet[21]比,提出的检测算法运行速度更快,且对中等目标和大目标的检测精度更高,更适用于这类目标的检测场景。图5展示了提出目标检测算法在PASCAL VOC 2007[18]和MS COCO[17]测试集的检测效果。

表1 提出的目标检测算法在MS COCO test-dev上的实验结果对比

Table 1 Experimental result comparison of proposed method to other methods on MS COCO test-dev

Method	Backbone	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	Time /ms
RetinaNet[22]	ResNet-101	39.1	59.1	42.3	21.8	42.7	50.2	90
FCOS[6]	ResNet-101	41.5	60.7	45.0	24.4	44.8	51.6	91
FSAF[5]	ResNet-101	40.9	61.5	44.0	24.0	44.2	51.3	138
ATSS[10]	ResNet-101	43.6	62.1	47.4	26.1	47.0	53.6	93
RepPoints[14]	ResNet-101	41.0	62.9	44.3	23.6	44.1	51.7	87
CornerNet[4]	Hourglass-104	40.5	56.5	43.1	19.4	42.7	53.9	227
ExtremeNet[7]	Hourglass-104	40.2	55.5	43.2	20.4	43.2	53.1	348
CenterNet[8]	Hourglass-104	42.1	61.1	45.9	24.1	45.5	52.8	298
LSNet[21]	ResNeXt-101	**43.9**	63.1	**47.8**	**26.6**	47.1	55.4	138
Proposed	ResNet-101	42.8	**65.1**	46.3	26.1	**47.3**	**55.8**	**78**

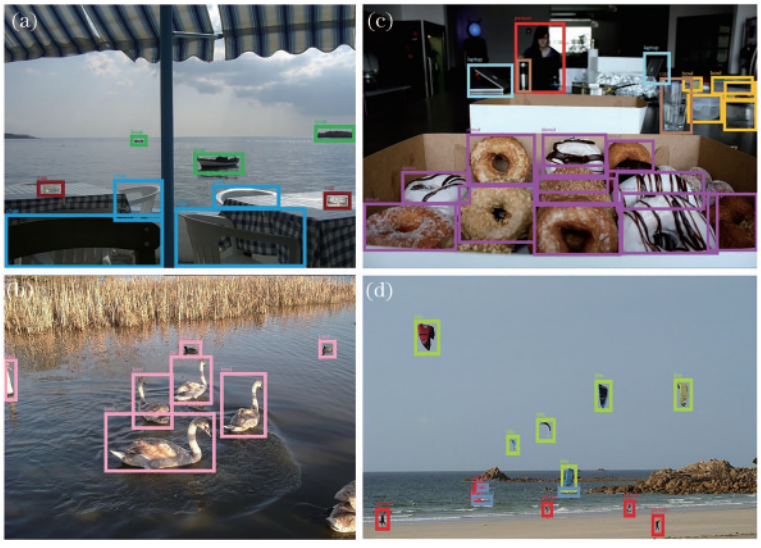

图5 提出检测算法在不同数据集的检测效果。(a)、(b) VOC 2007;(c)、(d) MS COCO

Fig. 5 Qualitative results of proposed method on different datasets. (a), (b) VOC 2007; (c), (d) MS COCO

3.3 消融实验

1) 模型设计。将RepPoints[14]作为基准网络,对所有模块的有效性进行了验证。在MS COCO的实验结果如表2所示,消融实验对比了在基准网络添加各模块后的检测器性能,其中LB表示基于语义的定位模块,FE为自适应特征增强模块,EO为高效的检测框优化模块。

基准网络RepPoints检测结果为39.1(AP),增加基于语义的定位模块,将检测精度大幅度提升到40.4(AP),对各个尺度目标的检测精度均有显著提高,说明基于语义的样本采样器能够更好地训练模型。增加特征增强模块后,总体检测精度提高到41.0(AP),检测精度较基准网络提高了约4%,证明利用语义信息和空间偏移进一步增强分类特征是有效的。最后,通过高效的检测框优化提升了检测精度,检测速度也提高了约10%。最终的目标检测器精度达到41.2(AP),检测速度约为77 ms,都得到了很大的提升。

2) 基于语义的定位模块。该模块中转换方法

表 2　将 3 个模块添加到 RepPoints 中的实验结果对比

Table 2　Experimental result comparison of integrating three modules into RepPoints

Method	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	Time /ms
Baseline RepPoints	39.1	60.5	42.1	21.2	41.2	50.0	85
RepPoints + LB	40.4	61.9	43.6	23.1	43.8	52.4	85
RepPoints + LB + FE	41.0	62.6	44.3	24.3	44.6	53.3	86
RepPoints + EO	39.2	61.7	42.2	21.4	42.8	50.6	76
Proposed (LB + FE+EO)	**41.2**	**63.0**	**44.2**	**24.6**	**45.0**	**53.9**	**77**

N_L: $F_R \rightarrow O_r$ 和阈值 ϵ_L 非常重要。表 3 表明 2.3 节提出的不同转换方法具有相似的检测结果。其中 max-pooling + avg-pooling 和 1×1 conv 方法取得了较高的检测精度，而 1×1 conv 可以达到更好的精度与速度的平衡，使检测器更专注于具有目标前景信息的像素点，抑制初始噪点。

表 3　不同转换方法的比较结果

Table 3　Comparison of different transformation conversions

Conversion function N_L	AP	AP_{50}	AP_{75}	Time /ms
1×1 conv	**40.4**	60.2	43.6	**85**
Max-pooling	40.3	60.1	43.4	**85**
Max-pooling + avg-pooling	**40.4**	**60.3**	**43.8**	86

阈值 ϵ_L 决定了样本采样器保留的特征点分布，采用不同的阈值时，采样器保留不同的正样本点。实验设置了不同的阈值，比较相应的检测结果。从表 4 可以看出，背景和语义信息较少的前景区域的概率值接近于 0，大量语义信息较少的负样本点可以通过一个较小的阈值过滤掉。

表 4　不同阈值 ϵ_L 的检测结果

Table 4　Detection results with different location thresholds

ϵ_L	AP	AP_{50}	AP_{75}
0.000	39.1	59.9	42.1
0.005	39.8	60.1	42.1
0.010	**40.4**	**60.2**	**43.6**
0.050	40.2	60.1	43.3
0.100	39.8	59.9	42.9

3）检测框优化方法。在不同的检测模型上应用两种检测框优化方法验证其高效性，实验结果如表 5 所示。相比级联式优化方法，本文提出的检测框优化方法具有更少的参数量（Param）和计算量（FLOPs）。但如 Chen 等[23]和 Ma 等[24]所述，参数量和计算量并不能充分反映卷积神经网络的运行效率，因此增加帧率（FPS）和运行时间作为进一步的评价指标。应用提出的检测框优化方法，运行速度大约提升 16%，检测精度也更高，实现了速度和精度的平衡。实验表明检测器的优异性能来自于更好的训练方法，而不是更多的参数或网络层，即使不采用可变形卷积进行第二阶段的检测框优化，也可以达到检测框优化的目的。

表 5　对不同网络应用高效检测框优化方法

Table 5　Impact of applying efficient refinement optimization to different networks

Method	EO	FLOPs /G	Param /M	FPS /(frame·s^{-1})	Time /ms	AP
RepPoints		190.16	36.62	11.8	85	39.1
	✓	**190.06**	**35.97**	**13.1**	**76**	**39.2**
Proposed		190.18	36.62	11.6	86	41.0
	✓	**190.07**	**35.97**	**12.9**	**77**	**41.2**

4　结　　论

针对无锚框目标检测器中存在的正、负样本划分不合理的问题，提出了基于语义的训练样本采样器，优先选择包含目标语义信息量较大的样本点为正样本；其次，为了使分类特征更加关注包含目标实例的区域，提出自适应特征增强模块，利用语义信息和定位偏移逐像素调整分类特征；最后，采用高效的检测框优化能够更好地训练模型，提升了检测速度，保证了特征的对齐性和一致性。基于以上模块，本文提出了一个单阶无锚框、点集表示目标的检测算法。消融和对照实验都表明该目标检测算法可以很好地实现检测速度和精度的平衡。在未来的工作中，该算法可以应用于其他基于中心的无锚框目标检测算法中，自适应特征增强模块和高效的检测框优化也普遍适用于目标检测任务，用以保持速度和精度的平衡。

参　考　文　献

[1] 孙迎春，潘树国，赵涛，等. 基于优化 YOLOv3 算法的交通灯检测[J]. 光学学报，2020，40(12)：1215001.
Sun Y C，Pan S G，Zhao T，et al. Traffic light detection based on optimized YOLOv3 algorithm[J]. Acta Optica Sinica，2020，40(12)：1215001.

[2] 薛芳芳，王月明，李琦. 基于特征部位空间关系的牛日常行为识别[J]. 激光与光电子学进展，2021，58(22)：2215007.
Xue F F，Wang Y M，Li Q. Recognition of cattle daily behavior based on spatial relationship of feature parts[J]. Laser & Optoelectronics Progress，2021，58(22)：2215007.

[3] 农元君，王俊杰. 基于嵌入式的遥感目标实时检测方法[J]. 光学学报，2021，41(10)：1028001.
Nong Y J，Wang J J. Real-time object detection in

remote sensing images based on embedded system[J]. Acta Optica Sinica, 2021, 41(10): 1028001.

[4] Law H, Deng J. CornerNet: detecting objects as paired keypoints[J]. International Journal of Computer Vision, 2020, 128(2): 642-656.

[5] Zhu C C, He Y H, Savvides M. Feature selective anchor-free module for single-shot object detection[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 840-849.

[6] Tian Z, Shen C H, Chen H, et al. FCOS: fully convolutional one-stage object detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 9626-9635.

[7] Zhou X Y, Zhuo J C, Krähenbühl P. Bottom-up object detection by grouping extreme and center points[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 850-859.

[8] Duan K W, Bai S, Xie L X, et al. CenterNet: keypoint triplets for object detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 6568-6577.

[9] Kong T, Sun F C, Liu H P, et al. FoveaBox: beyound anchor-based object detection[J]. IEEE Transactions on Image Processing, 2020, 29: 7389-7398.

[10] Zhang S F, Chi C, Yao Y Q, et al. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 9756-9765.

[11] Cai Z W, Vasconcelos N. Cascade R-CNN: delving into high quality object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 6154-6162.

[12] Li A, Yang X, Zhang C Y. Rethinking classification and localization for cascade R-CNN[EB/OL]. (2019-07-27)[2021-02-06]. https://arxiv.org/abs/1907.11914.

[13] Zhang S F, Wen L Y, Bian X, et al. Single-shot refinement neural network for object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 4203-4212.

[14] Yang Z, Liu S H, Hu H, et al. RepPoints: point set representation for object detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 9656-9665.

[15] Yang X, Yan J, Feng Z, et al. R3Det: refined single-stage detector with feature refinement for rotating object[C]// 2021 AAAI Conference on Artificial Intelligence, February 2-9, 2021, Vancouver, Canada, 35(4): 3163-3171.

[16] Kong T, Sun F C, Liu H P, et al. Consistent optimization for single-shot object detection[EB/OL]. (2019-01-19)[2021-05-07]. https://arxiv.org/abs/1901.06563v2.

[17] Lin T Y, Maire M, Belongie S J, et al. Microsoft COCO: common objects in context[M]//Fleet D, Pajdla T, Schiele B, et al. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8693: 740-755.

[18] Everingham M, Gool L, Williams C K I, et al. The pascal visual object classes (VOC) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[19] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 936-944.

[20] Chen K, Wang J Q, Pang J M, et al. MMDetection: open MMLab detection toolbox and benchmark[EB/OL]. (2019-06-17)[2021-04-05]. https://arxiv.org/abs/1906.07155v1.

[21] Duan K W, Xie L X, Qi H G, et al. Location-sensitive visual recognition with cross-IOU loss[EB/OL]. (2021-04-11)[2021-10-05]. https://arxiv.org/abs/2104.04899.

[22] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318-327.

[23] Chen Y H, Emer J, Sze V. Eyeriss: a spatial architecture for energy-efficient dataflow for convolutional neural networks[C]//2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture, June 18-22, 2016, Seoul, Korea (South). New York: IEEE Press, 2016: 367-379.

[24] Ma N N, Zhang X Y, Zheng H T, et al. ShuffleNet V2: practical guidelines for efficient CNN architecture design [M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11218: 122-138.