计算机科学
COMPUTER SCIENCE

基于多粒度的Transformer目标检测算法

徐放，苗夺谦，张红云

引用本文

徐放，苗夺谦，张红云. 基于多粒度的Transformer目标检测算法[J]. 计算机科学，2023，50(11)：143-150.

XU Fang, MIAO Duoqian, ZHANG Hongyun. Transformer Object Detection Algorithm Based on Multi-granularity [J]. Computer Science, 2023, 50(11): 143-150.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230600028

基于多粒度的 Transformer 目标检测算法

徐 放 苗夺谦 张红云
同济大学电子与信息工程学院 上海 201804
(2132959@tongji.edu.cn)

摘 要 与其他尺度目标不同,小目标具有携带语义信息和训练样本数量较少等特点。因此,当前目标检测算法存在小目标检测精度较低的问题。针对该问题,提出了基于多粒度的 Transformer 目标检测算法。首先,采用多粒度思想,设计了一种新的 Transformer 序列化方法,从粗到细逐个粒度地预测目标位置,从而提升模型的目标定位效果。然后,基于三支决策思想,细粒度挖掘小目标样本和常规尺度目标样本,从而增加小目标样本和难例负样本数量。最后,实验结果表明,在 COCO 数据集上,该算法的小目标检测精度(APs)达到了 31.5%,平均检测精度(mAP)达到了 49.1%;相比基线模型,AP_S 提升了 1.4%,mAP 提升了 2.2%;改进后的算法有效地提升了小目标检测效果,并显著提高了目标检测的整体精度。

关键词:小目标检测;多粒度;三支决策;Transformer;深度学习
中图法分类号 TP389.1

Transformer Object Detection Algorithm Based on Multi-granularity

XU Fang,MIAO Duoqian and ZHANG Hongyun
College of Electronic and Information Engineering,Tongji University,Shanghai 201804,China

Abstract Different from other scale objects,small objects have the characteristics of carrying less semantic information and a small number of training samples. Therefore,the current object detection algorithm has the problem of low detection accuracy for small objects. Aiming at this problem,a Transformer object detection algorithm based on multi-granularity is proposed. Firstly,adopting the multi-granularity idea,a new Transformer serialization method is designed to predict the object position granularly from coarse to fine,thereby improving the object location effect of the model. Then,based on the three-way decision idea,fine-grained mining of small object samples and regular-scale object samples increases the number of small object samples and hard-negative samples. Finally,experimental results on the COCO dataset show that,the small object detection average accuracy(APs) of the algorithm reaches 31.5%,and the mean average accuracy(mAP) reaches 49.1%. Compared with the baseline model,the APs is improved by 1.4% and the mAP is improved by 2.2%. The algorithm effectively improves the detection effect of small objects and significantly improves the overall accuracy of object detection.

Keywords Small object detection,Multi-granularity,Three-way decision,Transformer,Deep learning

1 引言

目标检测是计算机视觉领域的重点研究方向,被广泛应用于智能医疗、自动驾驶等多个领域。虽然深度学习技术的快速发展使得目标检测算法具备了较高的准确性,但小目标检测的精度仍然较低。在实际工业应用中,小目标在数据集中所占比例相对较高,因此对小目标检测的研究具有重要的实际意义。

小目标检测面临两个主要挑战。第一个挑战是小目标本身携带的语义信息较少,从而增加了模型在预测小目标位置时的困难。第二个挑战是在训练过程中,小目标的样本数量相对较少,限制了模型在小目标上的训练机会。因此,需要进一步改进小目标检测算法,以提升小目标检测的准确性。

针对目标定位困难的问题,基于卷积神经网络的检测方法通常采用预设锚框或根据特征图进行目标位置预测。然而,基于卷积神经网络的目标检测方法在处理尺度变化、

到稿日期:2023-06-03 返修日期:2023-09-18
基金项目:国家重点研发计划(2022YFB3104700);国家自然科学基金(62006172,61976158,61976160,62076182,62163016);江西省自然科学基金重点项目(20212ACB202001)
This work was supported by the National Key Research and Development Program of China(2022YFB3104700),National Natural Science Foundation of China(62006172,61976158,61976160,62076182,62163016) and Natural Science Foundation of Jiangxi Province,China(20212ACB202001).
通信作者:苗夺谦(dqmiao@tongji.edu.cn)

遮挡、复杂背景等问题时仍面临一定的困难。为了克服这些困难，近年来，部分方法通过引入 Transformer 的自注意力机制[1]，来提升模型的目标定位能力。然而，这些方法仅使用单一粒度表示目标位置，即仅使用目标边界框表示目标位置。对于携带语义信息较少的小目标，模型直接预测其目标边界框容易造成预测偏差。为了解决这一问题，本文采用多粒度思想[2-3]，以丰富目标位置信息的表示粒度。通过逐个粒度地预测目标位置，使得模型能够更准确地获取小目标位置信息，从而提高小目标检测的精度。

针对小目标样本数量较少的问题，目前主要有两种解决方法。第一种方法是通过修改损失函数结构，来增加小目标样本的权重，以便模型在训练过程中更加关注小目标。然而，这种方法虽然提高了小目标检测的效果，但未根本解决小目标样本数量较少的问题，仍存在模型忽视小目标的风险。第二种方法是采用数据增强策略，在训练数据中增加小目标的数量。然而，这种方法会增加额外的计算成本。为了解决这些问题，本文采用三支决策思想[4-6]，重新设计了样本划分方法，通过对样本空间进行细粒度挖掘，来增加小目标的样本数量，并且不会增加过多的计算量。通过这种方式，可以有效解决小目标样本数量较少的问题，从而提升小目标检测的性能。

本文的主要贡献如下：

1)提出了一个多粒度目标定位模块。该模块采用多粒度思想，从多个粒度层面定义目标位置信息，并基于 Transformer 架构逐个粒度地预测目标位置，从而实现全面精准的目标定位。

2)提出了一个三支决策样本划分模块。该模块采用三支决策思想，在常用样本划分方法丢弃的中间域中进行细粒度挖掘，确保在训练过程中有足够数量的小目标样本，从而提升小目标检测的性能。

3)在 COCO 数据集上进行对比实验和消融实验，以验证所提算法的有效性。相比基线模型，所提算法在 AP_S 上达到了 31.5%(提升了 1.4%)，在 mAP 上达到了 49.1%(提升了 2.2%)。

2 相关工作

2.1 目标检测方法

基于卷积神经网络[7](Convolutional Neural Network)的目标检测算法主要分为两类：基于锚框检测算法(Anchor-based)和无锚框检测算法(Anchor-free)。

基于锚框检测算法可分为双阶段检测算法(Two-stage)和单阶段检测算法(One-stage)。双阶段检测算法主要包括 R-CNN 系列算法[8-12]。该类方法分为两个阶段：第一阶段根据预设的锚框生成区域建议(Region Proposal)；第二阶段在区域建议的基础上预测结果。单阶段检测算法包含 SSD 算法[13]、RetinaNet 算法[14]和 YOLO 系列算法[15-18]，该类算法根据锚框直接预测目标位置和类别。其中，SSD 算法充分考虑了小目标所含语义信息较少的特殊性，该算法的思想是在细节信息丰富的浅层特征图上预测小目标。后续的多个小目标检测算法[19-20]是基于 SSD 算法的思想进行改进和发展。基于锚框的检测算法在预测小目标时存在一些问题，锚框尺度设计偏向于大目标，导致小目标检测性能较差，同时大量冗余的锚框增加了计算成本，降低了模型效率。

基于无锚框检测算法通过预测特征图上的目标关键点并进行关键点匹配来获取检测结果。该类算法包括：CornerNet[21]，FCOS[22]，CenterNet[23]，ExtremeNet[24]。无锚框的检测方法主要存在两个缺点。第一个缺点是由于小目标可利用的特征较少，导致模型无法准确预测小目标的关键点。第二个缺点是在关键点匹配的过程中，模型易受复杂背景的干扰，从而影响检测的准确性。本文采用 Transformer 机制和锚框机制相结合的方式，来提升小目标检测的精度，同时避免生成大量锚框。

2.2 基于 Transformer 的目标检测方法

鉴于 Transformer 模型在自然语言处理领域展现出了卓越的表现，越来越多的计算机视觉研究人员开始采用 Transformer 机制，并已取得重大突破[25-27]。DETR(Detection Transformer)[28]是首次采用 Transformer 机制的目标检测算法，该算法将目标检测问题转化为序列到序列的问题。DETR 的提出改变了目标检测思路，为目标检测领域的研究提供新的方向和启示。随后，许多学者对 DETR 进行改进。Zhu 等提出的 Deformable DETR[29]引入了可变形注意力机制，Liu 等提出的 DAB-DETR[30]将解码器可学习位置矩阵的语义定义为目标边界框。然而，在模型检测小目标时，由于小目标自身携带的语义信息较少，这些方法直接预测目标边界框通常存在预测偏差的问题。本文从多个粒度层面定义目标位置信息，并逐个粒度地预测目标位置，以避免直接预测细粒度目标位置，从而提升模型的目标定位效果。

3 基于多粒度的 Transformer 目标检测算法

本文提出的基于多粒度的 Transformer 目标检测算法包含多粒度目标定位模块(Multi-granularity Location，MGL)和三支决策样本划分模块(Three-way Decision Sample Division，TDSD)。算法的框架图如图 1 所示。

本文采用 ResNet-50[31]作为骨干网络，用于提取输入图片 $x_{\text{image}} \in R^{3*H_0*W_0}$ 的高质量特征。同时，利用 FPN[32]结构生成多尺度的特征图，并将高层特征的语义信息与低层特征的细节和位置信息相融合，以获取不同尺度的信息。

多粒度目标定位模块会从 5 个特征图中选取 k 个特征图进行预测，将选取的特征图逐个输入多粒度目标定位模块。为了充分利用不同尺度特征图之间细节信息和语义信息的差异，本文针对不同分辨率的特征图设置了不同尺寸的锚框，以便在具有丰富细节信息的特征图中进行小目标检测。将不同特征图上预测的细粒度目标位置信息统一缩放到分辨率最大的特征图上，进行分类预测和位置预测。

在训练阶段，本文采用三支决策样本划分模块，以增加小目标样本的数量，从而提升小目标的检测精度，并有效地挖掘难例负样本，从而提升模型的整体性能。

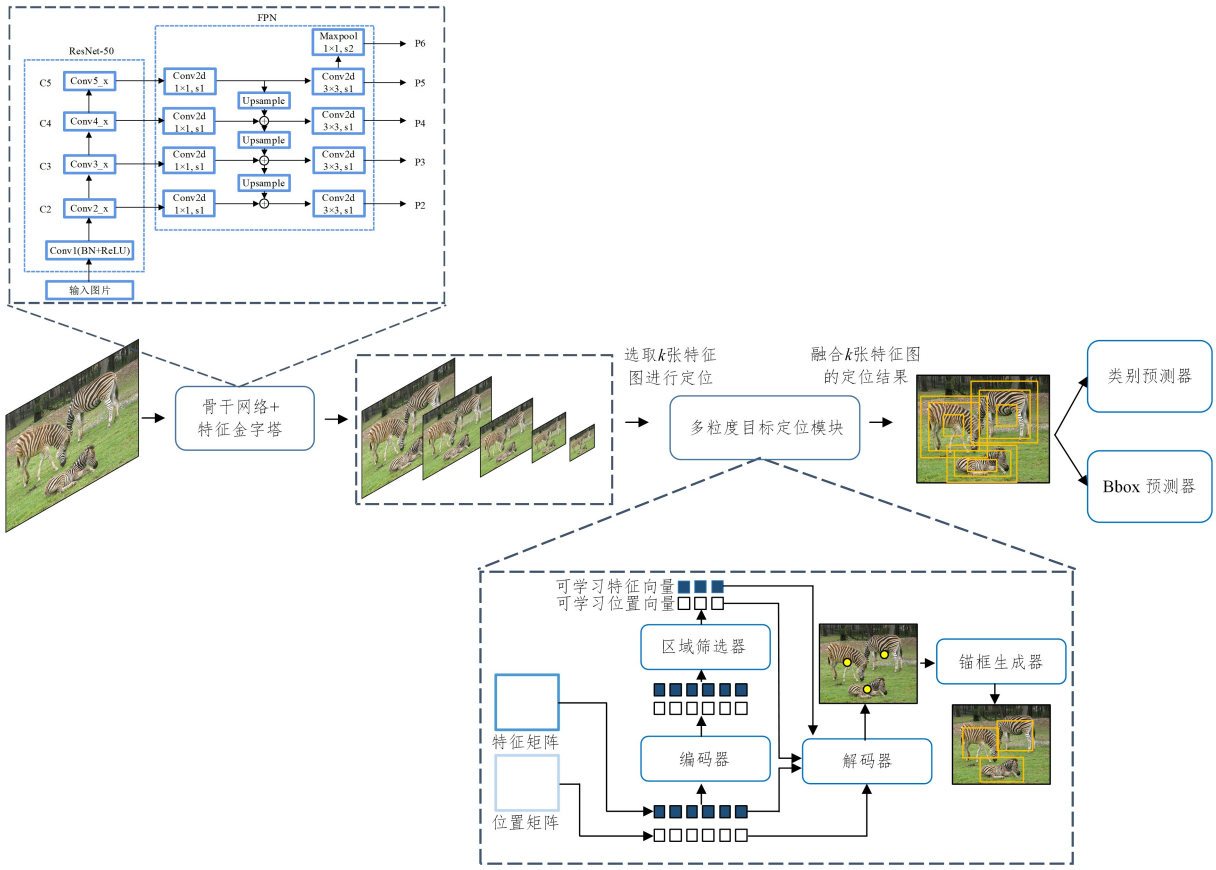


图 1　基于多粒度的 Transformer 目标检测算法框架

Fig. 1　Module framework for Transformer object detection algorithm based on multi-granularity

3.1　多粒度目标定位模块

Transformer 是一种基于自注意力机制的神经网络模型，其由编码器(Encoder)和解码器(Decoder)组成。编码器结构如图 2 所示，解码器结构如图 3 所示。注意力的计算式为：

$$Attention(\boldsymbol{Q},\boldsymbol{K})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{1}$$

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Attention(\boldsymbol{Q},\boldsymbol{K})\boldsymbol{V} \tag{2}$$

其中，$\boldsymbol{Q}$ 表示查询矩阵，$\boldsymbol{K}$ 表示键矩阵，$\boldsymbol{V}$ 表示值矩阵，d_k 为 $\boldsymbol{K}$ 矩阵的维度。通过式(1)计算注意力权重分配。注意力权重分配表示查询与键之间的相关性。根据注意力权重分配对值矩阵 $\boldsymbol{V}$ 进行加权求和操作，得到自注意力机制的输出矩阵。

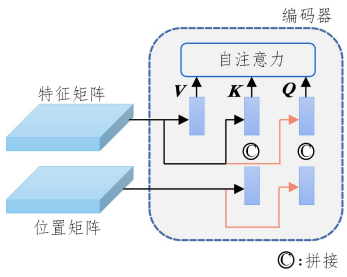


图 2　Transformer 编码器结构

Fig. 2　Transformer encoder structure

多粒度思想是一种认识和解决问题的思维方式，它将问题或数据分解为不同层次进行描述和分析。随着层数的加深，更深的层次蕴含更详细的特征，即粒度越来越细。多粒度思想从宏观到微观逐层剖析问题，通过综合分析不同层次的信息，来帮助形成全面准确的认知。在目标检测领域中，多尺度特征表示是多粒度思想的一种运用，它通过将特征图按不同尺度进行划分，有效地描述了目标在不同尺度下的特征信息。而本文提出的多粒度目标定位模块是根据目标位置信息进行粒度划分，并逐个粒度地预测目标位置，从而实现更准确的定位目标，有效提升目标定位的能力。

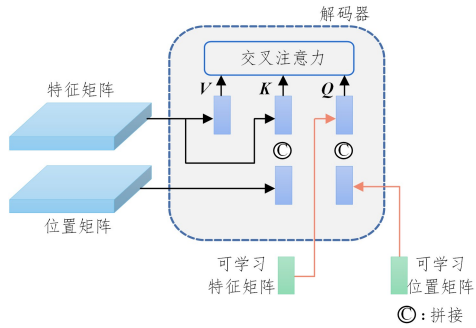


图 3　Transformer 解码器结构

Fig. 3　Transformer decoder structure

在目标检测任务中，模型通常采用四维向量(x,y,w,h)表示目标边界框，其中 x 和 y 分别表示目标中心的横坐标和纵坐标，w 和 h 分别表示目标边界框的宽和高。目标位置信息表示粒度单一，对于携带语义信息较少的小目标，直接利用其特征来预测详细的位置信息通常会出现偏差。

为了解决目标位置信息表示粒度单一的问题，本文基于多粒度思想，提出了多粒度目标位置定义，从多个粒度层面定义目标位置信息，包括粗粒度、中间层粒和度细粒度 3 种

粒度。在粗粒度层面，目标位置采用二维向量(x,y)表示。在中间层粒度层面，目标位置采用三维向量(x,y,r)表示，其中r表示目标的宽高比例。在细粒度层面，目标位置采用四维向量(x,y,w,h)表示。

基于多粒度目标位置定义，本文提出了一种新的 Transformer 序列化方法。该方法能够逐个粒度地获取目标的粗粒度位置信息与中间层粒度位置信息。一方面，该方法将 Transformer 解码器中可学习位置矩阵的语义定义为中间层粒度目标位置信息；另一方面，该方法重新设计了解码器结构，如图 4 所示。在预测携带语义信息较少的小目标时，该方法能够避免直接预测细粒度位置信息，从而解决 DAB-DETR 中存在的预测偏差问题。

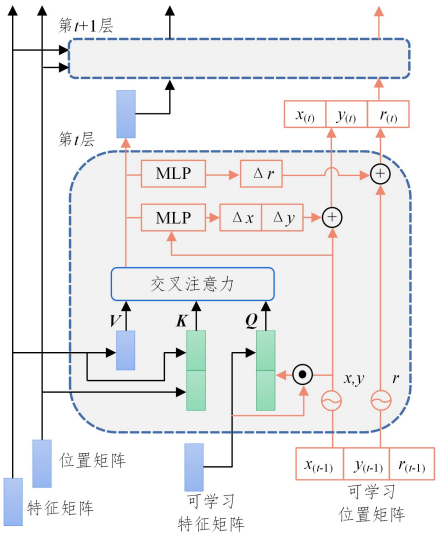


图 4 多粒度目标定位模块解码器结构

Fig. 4 Decoder structure of multi-granularity object localization module

多粒度目标定位模块结合了 Transformer 机制和锚框机制。首先，利用编码器进一步处理特征矩阵，设计一个区域筛选器，根据区域内存在目标的概率，筛选m个目标潜在区域。根据选定区域的特征向量和位置向量生成解码器的可学习特征矩阵$\boldsymbol{LF}$和可学习位置矩阵$\boldsymbol{LP}$。目标存在概率的计算式如下：

$$P_i=\mathrm{Softmax}(MLP(F_i)) \tag{3}$$

其中，P_i表示第i个区域目标存在概率，F_i表示第i个区域的特征，MLP 为多层感知机。解码器的可学习位置矩阵$\boldsymbol{LP}$和可学习特征矩阵$\boldsymbol{LF}$的具体定义如下：

$$\boldsymbol{LP}=Concat(\boldsymbol{pv}_1,\boldsymbol{pv}_2,\cdots,\boldsymbol{pv}_m) \tag{4}$$

$$\boldsymbol{LF}=Concat(\boldsymbol{pf}_1,\boldsymbol{pf}_2,\cdots,\boldsymbol{pf}_m) \tag{5}$$

其中，$\boldsymbol{pv}_i$表示目标存在概率为第i位的区域位置向量；$\boldsymbol{pf}_i$表示目标存在概率为第i位的区域特征向量。

然后，采用本文设计的交叉注意力机制逐层更新目标位置信息。在交叉注意力机制中，将可学习特征矩阵与可学习位置矩阵进行拼接，形成查询矩阵。同时，将特征矩阵和位置矩阵进行拼接，形成键矩阵。本文设计的交叉注意力机制利用查询矩阵和键矩阵的相关性调整权重分布，从而得到新的可学习特征矩阵。基于新的可学习特征矩阵，预测目标中心点偏移量$(\Delta x,\Delta y)$和目标宽高比例变化量Δr，从而完成对目标位置的更新，进一步提高位置信息的准确性。其计算式如下：

$$(\Delta x_t,\Delta y_t)=MLP(Concat(\boldsymbol{LF}_t,\boldsymbol{LP}_{t-1})) \tag{6}$$

$$\Delta r_t=MLP(\boldsymbol{LF}_t) \tag{7}$$

$$(x_t,y_t,r_t)=(x_{t-1}+\Delta x_t,y_{t-1}+\Delta y_t,r_{t-1}+\Delta r_t) \tag{8}$$

其中，x_t,y_t,r_t分别表示第t层解码器预测目标中心位置的横坐标、纵坐标、宽高比例；$\boldsymbol{LF}_t$表示第t层解码器交叉注意力机制输出的可学习特征矩阵；$\boldsymbol{LP}_{t-1}$表示第t层解码器交叉注意力机制输入的可学习位置矩阵。

最后，根据中间层粒度目标位置信息(x,y,r)和预设的锚框尺度，生成细粒度目标位置信息(x,y,w,h)。通过 Transformer 机制和锚框机制的结合，逐个粒度地获取更加精细的位置信息。这种方法确保在语义信息较少的情况下仍能进行精准的定位，进一步提升检测性能。

多粒度目标定位模块从 3 个粒度定义目标位置信息，并采用 Transformer 架构和锚框机制，逐个粒度地预测目标位置信息。该模块解决了模型直接预测目标边界框时出现偏差的问题，从而在检测携带语义信息较少的小目标时，可以实现精准的位置预测。

3.2 三支决策样本划分模块

常用的样本划分策略设定两个重叠度（Intersection over Union，IoU）阈值$thre_{\mathrm{pos}}$和$thre_{\mathrm{neg}}$。根据目标预测框和目标真实框之间的重叠度划分正负样本。当样本重叠度大于$thre_{\mathrm{pos}}$时，将其划分为正样本；当样本重叠度小于$thre_{\mathrm{neg}}$时，将其划分为负样本；当样本重叠度介于$thre_{\mathrm{pos}}$和$thre_{\mathrm{neg}}$之间时，将其丢弃。具体的样本划分方式如下：

$$P_i^*=\begin{cases}\text{positive sample}, & IoU_i>thre_{\mathrm{pos}}\\ \text{discarded sample}, & thre_{\mathrm{neg}}\leqslant IoU_i\leqslant thre_{\mathrm{pos}}\\ \text{negtive sample}, & IoU_i<thre_{\mathrm{neg}}\end{cases} \tag{9}$$

其中，P_i^*表示第i个样本类别；IoU_i表示第i个样本的重叠度。

由于小目标的分辨率较小，小目标的定位准确度要求更高，当检测模型产生预测偏差时，小目标的重叠度相对较低。如图 5 所示，其中小目标分辨率为 50×50，大目标分辨率为 200×200。

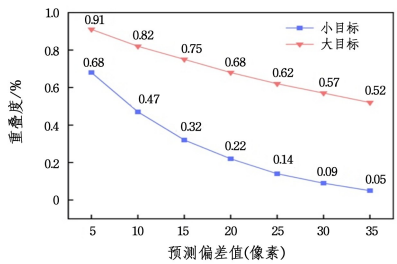


图 5 不同尺度目标预测误差的影响

Fig. 5 Prediction deviation influence of different scale objects

基于这种现象，常用的样本划分方式通常无法获取足够数量的小目标样本。除此之外，负样本分为易例负样本（Simple Negative Sample）和难例负样本（Hard Negative Sample）。难例负样本指难以预测其类别的样本。与其他负样本相比，难例负样本对提升模型性能具有更高的参考价值。然而，当前的样本划分方法采用随机抽取负样本的方法，导致难例负样本数量较少。

为了解决这些问题，本文提出了三支决策样本划分模块，该模块细粒度地挖掘被抛弃的中间域样本空间，旨在确保小目标样本和难例负样本在数据集中占据适当比例。该模块的具体框架如图 6 所示。

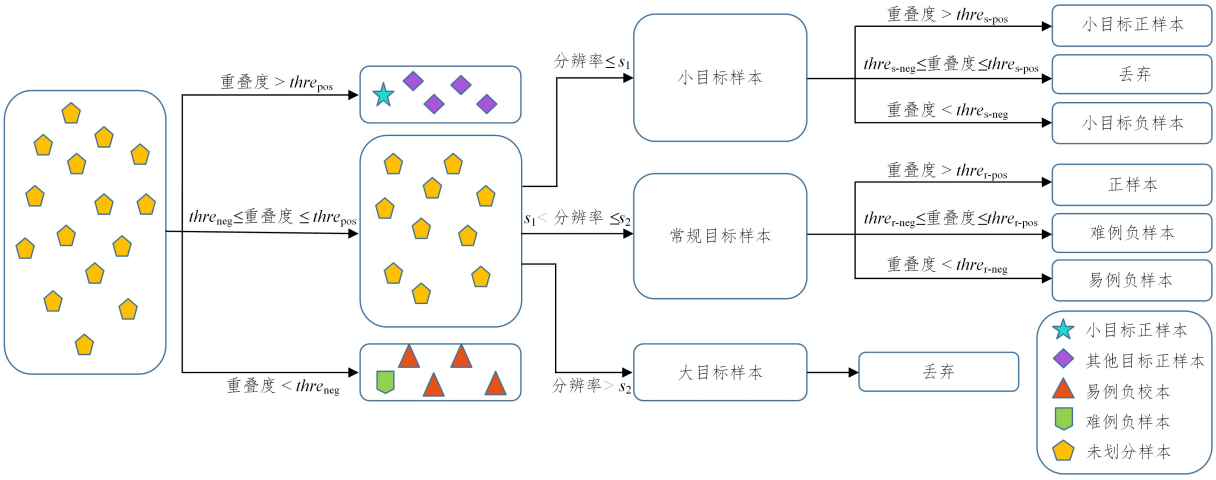


图 6　三支决策样本划分模块框架

Fig. 6　Module framework for three-way decision sample division

本文参考 COCO 数据集中对不同尺度目标的定义，设定两个分辨率阈值 s_1 和 s_2。其中 s_1 为 32×32，s_2 为 95×95。根据目标样本的分辨率大小，对中间域样本进行分类。当目标样本的分辨率小于 s_1 时，将其归为小目标样本；当目标样本分辨率介于 s_1 和 s_2 之间时，将其归为常规尺度目标样本；当目标样本分辨率大于 s_2 时，将其归为大目标样本。具体划分方式如下：

$$S_{\text{mid-area}}=\begin{cases}\text{Big Object}, & s_i>s_2\\ \text{Regular Object}, & s_1<s_i\leqslant s_2\\ \text{Small Object}, & s_i\leqslant s_1\end{cases}\tag{10}$$

其中，$S_{\text{mid-area}}$ 表示在采用常用样本划分方法时处于中间域的样本，s_i 表示第 i 个样本的分辨率。

大目标具有较丰富的语义信息，当前的目标检测模型通常能够准确地检测大目标。因此，本文直接丢弃处于中间域的大目标样本。

常规目标样本数量在数据集中占据很大比例，是获取难例负样本的主要来源。本文针对常规目标样本设置了一对重叠度正负阈值 $thre_{\text{r-pos}}$ 和 $thre_{\text{r-neg}}$。当样本重叠度大于正阈值 $thre_{\text{r-pos}}$ 时，将其划分为正样本，当样本重叠度小于负阈值 $thre_{\text{r-neg}}$ 时，将其划分为易例负样本，而当样本重叠度处于正负阈值中间时，将其划分为难例负样本。

针对小目标样本，本文设置了一对更加适合小目标的重叠度正负阈值 $thre_{\text{s-pos}}$ 和 $thre_{\text{s-neg}}$。当样本重叠度大于正阈值 $thre_{\text{s-pos}}$ 时，将其划分为正样本，当样本重叠度小于 $thre_{\text{r-neg}}$ 时，将其划分为负样本，当样本重叠度处于正负阈值中间时，其被丢弃处理。

三支决策样本划分模块通过对样本空间进行细粒度的挖掘，增加了小目标样本和难例负样本的数量，从根本上解决了样本数量不足的问题，并没有引入过多的计算量。

3.3　损失函数

本文算法的损失由分类损失和边界框回归损失构成。

本文的分类损失函数是当前目标检测领域通用的多分类交叉熵损失函数，具体表达式如下：

$$\mathcal{L}_{\text{cls}}=-\sum_{i=1}^{C}y_i\log(p_i)\tag{11}$$

其中，$\boldsymbol{y}_i$ 表示真实类别向量，真实类别对应的元素值为 1，其他类别的元素值均为 0；P_i 为预测对应类别的概率；C 表示样本数量。

本文模型检测过程中涉及目标的宽高比例 r 的预测。为了考虑宽高比例在回归损失计算中的影响，本文采用的边界框回归损失函数是 CIoU 损失函数[33]。具体边界框回归损失如下：

$$\mathcal{L}_{\text{bbox}}=\mathcal{L}_{\text{CIoU}}=1-CIoU\tag{12}$$

$$CIoU=1-IoU+\frac{b^2}{c^2}+\alpha\nu\tag{13}$$

$$\nu=\left(\arctan\frac{w^{gt}}{h^{gt}}-\arctan\left(\frac{w}{h}\right)\right)^2\tag{14}$$

其中，b 表示目标真实框中心到预测框中心的距离；c 表示目标真实框和预测框最小闭包矩形的对角线长度；α 为常量值，根据以往工作[33]的建议将其设置为 $\frac{4}{\pi^2}$；ν 保证了损失函数能够有效地考虑目标的宽高比例因素。

本文算法的整体损失函数如下：

$$\mathcal{L}=\lambda_1\mathcal{L}_{\text{cls}}+\lambda_2L_{\text{bbox}}\tag{15}$$

其中，λ_1 和 λ_2 为常量，本文中 λ_1 和 λ_2 统一设置为 0.5，以平衡两个损失的贡献。

4　实验结果

4.1　数据集和评价指标

COCO 数据集[34]是目标检测领域最具挑战性的数据集之一，其中包含约 33 万张图片和 150 万个目标实例，涵盖了 80 种不同的目标类别。同时，该数据集从真实世界中采集，包含了不同的场景和环境，例如室内和室外、城市街道、公园、办公室、野外等多种场景。其中，COCO 数据小目标占比达到 41.43%，并且本文的研究目标是改进小目标检测，因此本文选用 COCO 数据集作为本次实验的数据集。

本文采用目标检测领域通用的评估指标——平均精度（mean Average Precision，mAP）作为衡量指标来评估实验

结果。除了 mAP，还使用其他指标详细地评估算法的性能。其中，AP_S，AP_M 和 AP_L 分别表示小、常规、大目标的平均检测准确率；AP_{50} 和 AP_{75} 分别表示重叠度阈值设置为 0.5 和 0.75 时的准确率。通过这些指标，能够全面评估本文算法在小、常规、大目标检测方面的准确度，并提供更多的信息来衡量模型的性能。

4.2 训练设置

本文使用 PyTorch 框架来实现本文提出的算法。在实验中，采用 ResNet-50 作为骨干网络，并且在 ImageNet[35] 数据集上进行预训练。在训练过程中，采用随机梯度下降算法作为模型优化器，将动量参数设置为 0.9，权重衰弱参数设置为 0.0001，批量大小设置为 8。本文参考以往工作[9-12]，并从对比实验公平性的角度，根据常用样本划分方法的重叠度正负阈值进行设置。具体而言，本文将正阈值 $thre_{pos}$ 设置为 0.7，将负阈值 $thre_{neg}$ 设置为 0.3。

4.3 算法参数分析

本文通过大量实验分析多粒度目标定位模块的参数设置。具体参数包括选取特征矩阵个数 k、特征图的选取结果、针对不同特征图所设定的锚框边长，以及区域筛选器输出的区域个数 m。具体实验结果如表 1 所列。根据实验结果，多粒度目标定位模块参数依据表 1 中的第 3 行进行设置。

表 1 多粒度目标定位模块参数设置与实验结果

Table 1 Parameter setting and experimental results of multi-granularity location module

k	选层结果	锚框边长	m	AP	AP_S	AP_M	AP_L
3	$P_2P_4P_6$	64,238,424	100	47.2	30.2	50.4	62.4
4	$P_2P_3P_4P_5$	54,153,285,426	80	47.7	30.6	51.0	62.7
5	$P_2P_3P_4P_5P_6$	42,136,248,342,436	70	48.3	30.9	51.5	63.1

本文针对三支决策样本划分模块参数设定进行大量实验。根据 COCO 数据集对小目标和常规尺度目标的分辨率进行定义，设定分辨率阈值参数 s_1 为 32×32，参数 s_2 为 95×95。实验结果如表 2 所列。对实验结果进行分析，依据表 2 第 4 行设置小目标正负阈值 $thre_{s\text{-}pos}$ 和 $thre_{s\text{-}neg}$，常规目标正负阈值 $thre_{r\text{-}pos}$ 和 $thre_{r\text{-}neg}$。

表 2 三支决策样本划分模块参数设置与实验结果

Table 2 Parameter settings and experimental results of three-way decision sample division module

$thre_{s\text{-}pos}$	$thre_{s\text{-}neg}$	$thre_{r\text{-}pos}$	$thre_{r\text{-}neg}$	AP	AP_S	AP_M	AP_L
0.65	0.50	0.60	0.5	48.4	31.0	51.5	63.0
0.65	0.50	0.65	0.4	48.7	31.1	51.7	63.4
0.60	0.45	0.60	0.5	48.6	31.3	51.4	63.1
0.60	0.45	0.65	0.4	49.1	31.5	52.2	63.7

4.4 对比实验

本文在 COCO 数据集上进行对比实验，评估了不同目标检测模型的性能。这些模型包含基于锚框的目标检测算法、无锚框目标检测算法和采用 Transformer 机制的目标检测算法。对比实验结果如表 3 所列。

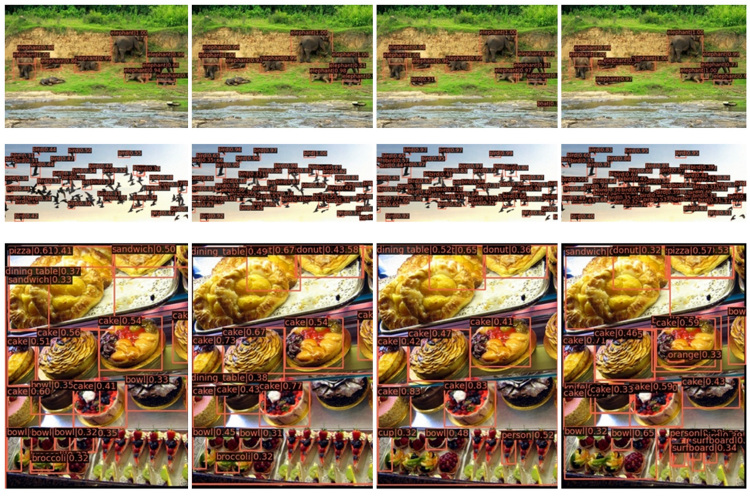
表 3 对比实验结果

Table 3 Comparison of experimental results

模型	骨干网络	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster R-CNN[10]	ResNet101	36.2	59.1	39.0	18.2	39.0	48.2
Mask R-CNN[11]	ResNet101	38.2	60.3	41.7	20.1	41.1	50.2
Cascade R-CNN[12]	ResNet101	42.8	62.1	46.3	23.7	45.5	55.2
RetinaNet[14]	ResNet101	39.1	59.1	42.3	21.8	42.7	50.2
YOLOv3[16]	Darknet53	33.0	57.9	34.4	18.3	35.4	41.9
SSD[13]	ResNet101	31.2	50.4	33.3	10.2	34.5	49.8
CornetNet[21]	Hourglass104	42.1	57.8	45.3	20.8	44.8	56.7
ExtremeNet[24]	Hourglass104	43.7	60.5	47.0	24.1	46.9	57.6
DETR[28]	ResNet50	43.3	63.1	45.9	22.5	47.3	61.1
Deformable DETR[29]	ResNet50	46.2	65.2	50.0	28.8	49.2	61.7
DAB-DETR[30]	ResNet50	46.9	66.0	50.8	30.1	50.4	62.5
Ours	ResNet50	**49.1**	67.2	**52.7**	**31.5**	**52.2**	63.7

根据对比实验结果可知，本文算法在 mAP 上达到了 49.1%，在 AP_S 上达到了 31.5%。与基线模型 DAB-DETR 相比，mAP 提升了 2.2%，AP_S 提升 1.4%。此外，与其他目标检测模型相比，本文算法在多个评价指标上展现出了明显的优势。

为了更加直观地展示本文算法的有效性和健壮性，本文展示了不同检测模型在不同场景下的可视化结果。图 7 给出了 3 组不同的可视化结果对比。



(a)Faster R-CNN (b)CornerNet (c)DAB-DETR (d)OURS

注：第 1 行图片为室外场景，第 2 行图片为小目标密集分布场景，第 3 行图片为室内场景。

图 7 不同场景的可视化结果对比

Fig. 7 Visualization results comparison of different scenes

3 组图片涵盖了室内和室外等不同场景，其中第二组图片涉及小目标密集分布的复杂场景。每组检测采用 4 种不同检测算法，从左到右依次是：Faster R-CNN、CornerNet、DAB-DETR 以及本文算法。对比同一组图片的检测结果发现，本文算法可以更精准地预测出小目标，整体的检测效果优于其他算法，验证了本文算法的有效性。对比不同图片的检测结果发现，在所有检测场景中，本文算法均优于其他算法，并没有因场景变化而影响检测效果，验证了本文算法在多个检测场景中的健壮性。

4.5　消融实验

为了更好地验证本文提出模块的有效性，分别对多粒度目标定位模块（MGL）和三支决策样本划分模块（TDSD）进行消融实验。消融实验结果如表 4 所列。

表 4　消融实验结果

Table 4　Results of ablation experiment

MGL	TDSD	mAP	AP_S
		46.9	30.1
√		48.3	30.9
√	√	49.1	31.5

由于基线模型 DAB-DETR 采用匈牙利算法完成正负样本划分，而不是根据重叠度完成样本划分，因此无法直接在基线模型上采用三支决策样本划分模块。对比表 4 的第 1 行和第 2 行可以发现，添加多粒度目标定位模块（MGL）后中，mAP 提升了 1.4%并且 AP_S 提升了 0.8%。这一结果表明，通过逐个粒度地进行目标位置预测，可以有效提高模型的定位性能，解决了在语义信息不明确的情况下出现的预测偏差问题。通过对比表 4 的第 2 行和第 3 行发现，在引入三支决策样本划分模块（TDSD）后，mAP 提升了 0.8%，AP_S 提升了0.6%。这表明，通过对样本空间进行细粒度的挖掘，有效地增加了小目标样本和难例负样本的数量，从而提升了检测精度。

结束语　针对小目标定位效果差的问题，本文提出了多粒度目标定位模块，从多个粒度定义目标位置信息，并采用逐个粒度预测目标位置的方式，解决了小目标直接预测目标边界框导致的误差问题。针对小目标训练样本较少的问题，本文提出了三支决策样本划分模块，对样本空间进行细粒度挖掘，从而增加小目标样本的数量。实验结果证明，本文方法明显提升了小目标检测的性能，同时提高了目标检测的整体精度。然而，本文引入了较多的参数，导致模型推理速度较慢。下一步将继续研究在不引入过多参数的前提下，显著地提高模型的检测性能。

参 考 文 献

[1] VASWANI A, SHAZEERN, PARMAR N, et al. Attention Is All You Need[C]// Advances in Neural Information Processing Systems. Curran, 2017: 5998-6008.

[2] WANG Z Y, MIAO D Q, ZHAO C R, et al. A Pedestrian Tracking Algorithm Based on Multi-Granularity Feature[J]. Journal of Computer Research and Development, 2020, 57(5): 996-1002.

[3] CHEN Y F, MIAO D Q. Granular Regression with A Gradient Descent Method[J]. Information Sciences, 2020, 537: 247-260.

[4] QIAN J, LIU C H, MIAO D Q, et al. Sequential Three-way Decisions via Multi-granularity [J]. Information Sciences, 2020, 507: 606-629.

[5] YUE X D, CHEN Y F, MIAO D Q, et al. Fuzzy Neighborhood Covering for Three-way Classification[J]. Information Sciences, 2020, 507: 795-808.

[6] LANGG M, MIAOD Q, HAMIDO F. Three-way Group Conflict Analysis Based on Pythagorean Fuzzy Set Theory[J]. IEEE Transactions on Fuzzy Systems, 2020, 28(3): 447-461.

[7] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet Classification with Deep Convolutional Neural Networks[J]. Advances in Neural Information Processing Systems, 2012, 25: 1097-1105.

[8] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation [C] // Computer Vision and Pattern Recognition (CVPR). NJ: IEEE, 2014: 1714-1722.

[9] GIRSHICK R. Fast R-CNN[C]// International Conference on Computer Vision(ICCV). Cham: Springer, 2015: 1440-1448.

[10] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.

[11] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]// International Conference on Computer Vision (ICCV). Cham: Springer, 2017: 2980-2988.

[12] CAI Z W, VASCONCELOS N. Cascade R-CNN: Delving into High-Quality Object Detection[C]// Computer Vision and Pattern Recognition(CVPR). NJ: IEEE, 2018: 6154-6162.

[13] WEI L, DRAGOMIR A, DUMITRU E, et al. SSD: Single Shot MultiBox Detector[C]// European Conference on Computer Vision(ECCV). Cham: Springer, 2016: 21-37.

[14] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal Loss for Dense Object Detection[C]// Conference on Computer Vision(ICCV). Cham: Springer, 2017: 2980-2988.

[15] REDMON J, FARHADI A. YOLO9000: Better, Faster, Stronger [C]// IEEE Conference on Computer Vision and Pattern Recognition(CVPR). NJ: IEEE, 2017: 6517-6525.

[16] REDMON J, FARHADI A. YOLOv3: An Incremental Improvement[J]. arXiv: 1804.02767, 2018.

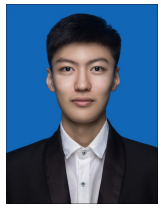
[17] BOCHKOVSKIY A, WANG C, LIAO H. YOLOv4: Optimal Speed and Accuracy of Object Detection[J]. arXiv: 2004.10934, 2020.

[18] WANG C, BOCHKOVSKIY A, LIAO H. YOLOv7: Trainable Bag-of-freebies Sets New State-of-the-art for Real-time Object Detectors[J]. arXiv: 2207.02696, 2022.

[19] ZHANG W L, CHEN X H. SSD Object Detection Algorithm with Cross-layer Fusion and Receptive Field Amplification[J]. Computer Science, 2023, 50(3): 231-237.

[20] JIA T H, PENG L. SSD Object Detection Algorithm with Residual Learning and Cyclic Attention[J]. Computer Science, 2023, 50(5): 170-176.

[21] LAW H, DENG J. CornerNet: Detecting Objects as Paired Keypoints[C]// European Conference on ComputerVision(ECCV). Cham: Springer, 2018.
[22] TIAN Z, SHEN C H, CHENH, et al. FCOS: Fully Convolutional One-Stage Object Detection[C]// International Conference on Computer Vision(ICCV). NJ: IEEE, 2019: 9627-9636.
[23] DUAN K W, BAI S, XIE L X, et al. CenterNet: Keypoint Triplets for Object Detection[C]// IEEE/CVF International Conference on Computer Vision(ICCV). Cham: Springer, 2019: 1-16.
[24] ZHOU X Y, ZHUO J C, PHILIPP K. Bottom-Up Object Detection by Grouping Extreme and Center Points[C]// Conference on Computer Vision and Pattern Recognition(CVPR). NJ: IEEE, 2019: 850-859.
[25] LIU Z, LIN Y T, CAO Y, et al. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). NJ: IEEE, 2021: 10012-10022.
[26] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is Worth 16x16 Words: Transformers for Image Recognition at Scale[J]. arXiv: 2010. 11929, 2020.
[27] WANG W H, XIE E Z, LI X, et al. PyramidVision Transformer: A Versatile Backbone for Dense Prediction without Convolutions[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). NJ: IEEE, 2021: 13977-13986.
[28] NICOLAS C, FRANCISCO M, GABRIEL S, et al. End-to-end Object Detection with Transformers[C]// European Conference on Computer Vision(ECCV). Cham: Springer, 2020: 586-603.
[29] ZHU X Z, SU W J, LU L W, et al. Deformable DETR: Deformable Transformers for End-to-End Object Detection[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). NJ: IEEE, 2021.
[30] LIU S L, LI F, ZHANG H, et al. DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR[J]. arXiv: 2201. 12329, 2022.
[31] HE K, ZHANG X Y, REN S Q, et al. Deep Residual Learning for Image Recognition[C]// Computer Vision and Pattern Recognition(CVPR). NJ: IEEE, 2016: 770-778.
[32] LIN T Y, DOLLÁR P, GIRSHICK R B, et al. Feature Pyramid Networks for Object Detection[C]// Computer Vision and Pattern Recognition(CVPR). NJ: IEEE, 2017: 2117-2125.
[33] ZHANG Z, QI H, LIU S, et al. CIoU: Enhancing Convolutional Neural Networks for Object Detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV). NJ: IEEE, 2019: 7155-7163.
[34] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common Objects in Context[C]// European Conference on Computer Vision(ECCV). Cham: Springer, 2014: 740-755.
[35] DENG J, WEI D, SOCHER R, et al. ImageNet: A Large-scale Hierarchical Image Database[C]// Computer Vision and Pattern Recognition. Florida(CVPR), NJ: IEEE, 2009: 248-255.



XU Fang, born in 1999, postgraduate. His main research interests include object detection and granular computing.

MIAO Duoqian, born in 1964, professor, Ph.D supervisor. His main research interests include rough set and machine learning.

(责任编辑:喻藜)