doi:10.3969/j.issn.1673-9833.2023.01.009

基于注意力机制和多层次特征融合的目标检测算法

周秋艳[1]，肖满生[1]，范双南[2]

（1. 湖南工业大学 计算机学院，湖南 株洲 412007；2. 湖南交通工程学院 电气与信息工程学院，湖南 衡阳 421001）

摘　要：为了提高目标检测的准确率，提出一种基于注意力机制和多层次特征融合的图像目标检测算法。该算法在 Cascade R-CNN 模型的基础上，以 RseNet50 为主干网络，通过嵌入简单的注意力模块（SAM）来提高网络的判别能力；其次，利用深度可分离卷积改进特征金字塔网络（FPN），设计了多层次特征融合模块（MFFM），对多尺度特征进行融合，以丰富特征图的信息量，并对不同层次的特征图赋予相应的权重以平衡不同尺度的特征信息；最后，结合目标检测方法中的区域建议网络（RPN）结构获取目标的候选区域进行分类和回归处理，确定检测目标的位置和类别。实验结果表明，相较于 Cascade R-CNN 目标检测算法，该算法的检测精度提升了约 2.0%。

关键词：目标检测；注意力模块；多层次特征融合；深度可分离卷积

中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1673-9833(2023)01-0061-08

引文格式：周秋艳，肖满生，范双南 . 基于注意力机制和多层次特征融合的目标检测算法 [J]. 湖南工业大学学报，2023，37(1)：61-68.

Target Detection Algorithm Based on Attention Mechanism and Multi-Level Feature Fusion

ZHOU Qiuyan[1]，XIAO Mansheng[1]，FAN Shuangnan[2]

（1. College of Computer Science，Hunan University of Technology，Zhuzhou Hunan 412007，China；
2. College of Electrial and Information Engineering，Hunan Institute of Transportation Engineering，Hengyang Hunan 421001，China）

Abstract：For an improvement of the accuracy of object detection, an image object detection algorithm has thus been proposed based on attention mechanism and multiple feature fusion. On the basis of the Cascade R-CNN model, the algorithm uses RseNet50 as the backbone network, with a simple attention module (SAM) embedded so as to improve the discrimination ability of the network. Secondly, a multi-level feature fusion module (MFFM) is designed by using the deep separable convolution to improve the feature pyramid network (FPN), followed by a fusion of the multi-scale features to enrich the information of feature maps, with the corresponding weights given to the feature maps of different levels to balance the feature information of different scales. Finally, combined with the region proposal network (RPN) structure in the target detection method, the candidate regions of the target can be obtained for classification and regression processing to determine the location and category of the detection target. Experimental results show that compared with Cascade R-CNN target detection algorithm, the detection accuracy of the proposed

收稿日期：2022-03-16

基金项目：湖南省自然科学基金资助项目（2021JJ50049）；湖南省教育厅科学研究基金资助重点项目（21A0607）

作者简介：周秋艳（1997-），女，云南曲靖人，湖南工业大学硕士生，主要研究方向为深度学习，
E-mail：zhouqy014@foxmail.com

通信作者：肖满生（1968-），男，湖南株洲人，湖南工业大学教授，硕士生导师，主要研究方向为智能计算和智能信息处理，
E-mail：349407041@qq.com

algorithm has been improved by approximately 2.0%.

Keywords：target detection；attention mechanism；multi-level feature fusion；deep separable convolution

1 研究背景

目标检测是指从图像中获取感兴趣的目标，确定每个目标的准确位置和类别，并在图像上进行标注。近年来随着目标检测的快速发展，该技术被广泛应用于智能驾驶、医学图像诊断、行人检测和航天航空等领域[1−3]。

基于手工特征提取的传统检测算法主要包括以下步骤：图像预处理、窗口滑动、特征提取和特征数据处理、分类器分类[4]。这些算法在特征提取阶段常用的视觉特征有 Harr 特征[5]、HOG 特征[6]、SIFT（scale-invariant feature transform）特征[7]等，但这些特征被用于识别特定的任务时往往存在一些缺陷。如依靠人工的先验知识设计特征提取器，缺乏一定客观性，因此对多样性目标的检测鲁棒性差，在复杂场景下很难取得较好的效果，检测精度和速度较低。近年随着深度学习的迅速发展，许多学者利用深度卷积神经网络（convolutional neural networks，CNN）进行特征提取，该模型的泛化能力较强，目标检测精度和速度得到了较大提升。目前主流的目标检测算法主要分为单阶段和两阶段两种策略，基于候选框的两阶段方法如 R-CNN（region-based convolutional neural networks）[8]、Faster RCNN[9]、Cascade RCNN[10]等，其实现过程为：先对感兴趣的区域进行候选框获取，而后利用 CNN 网络生成对应的特征图，对候选框进行分类识别和边框回归，完成目标检测，此类方法检测精度较高，但实时性不强。而基于回归的单阶段方法如 SSD（single shot multibox detector）[11] 和 YOLO（you only look once）[12] 等，此类方法利用 CNN 网络直接预测目标的类别与位置，无需获取候选框，检测的实时性较强，但精度不高。针对这些问题，专家学者提出了许多基于深度学习框架模型以改善目标检测效果，对于不同尺度的目标需要不同大小感受野的特征去识别，而神经网络的高层特征中包含了丰富的语义信息，因此许多方法是通过增加网络层数来获得语义信息更强的高层特征图，从而提升网络性能，但随着卷积层数增加，图像经过大量特征处理后，目标的位置信息变弱。高层特征图的语义信息较强、分辨率较低，而低层特征图的分辨率较高、语义信息较弱，同时相邻层级的特征图间的相关性在此过程中会愈加弱化，导致分类和回归的精度不高。针对这些问题，一系列典型的多尺度特征融合模块被提出，如特征金字塔网络（feature pyramid network，FPN）[13]、神经结构搜索特征金字塔网络（neural architecture Search Feature Pyramid Network，NAS-FPN）[14]，以及许多运用了多尺度特征融合方法的网络：如 PANet（path aggregation network）[15]、HRNet（high-resolution representation learning for human pose estimation）[16]。Tan M. X. 等[17] 提出了一种加权双向特征金字塔网络（bi-directional feature pyramid network，Bi-FPN）实现快速的多尺度特征融合。Cao J. X. 等[18] 通过整合注意力引导的多路径特征，利用了来自各种大感受野的判别信息，提出注意力引导的上下文特征金字塔网络。Xing H. J. 等[19] 提出了基于双重注意力机制的特征金字塔网络，改善了小目标检测效果，Hu J. 等[20] 对通道之间的相互依赖性进行建模以自适应地重新校准通道特征响应，提出了 SENet（squeeze and excitation networks），极大改善了网络性能。

受上述思想启发，本文提出了一种基于注意力机制和多层次特征融合的目标检测算法，能够有效提高目标检测精度，主要贡献如下：

1）设计了简单的注意力模块（simple attention module，SAM），并将其应用于主干网络，对网络通道关系进行建模以增强网络的表征能力；

2）针对检测中的多尺度问题及网络中不同分辨率的特征对网络性能提升贡献的不同，本文设计了基于深度可分离卷积的多层次特征融合模块（multi-layer feature fusion module，MFFM），对多尺度特征进行融合，在保证效率的情况下丰富了特征信息，同时引入可学习的权重，获取不同输入特征的重要性程度，以更好地平衡不同尺度的特征信息。

2 目标检测算法

2.1 简单的注意力模块

注意力机制是在全局信息中获得需要关注的部分的一种方式。本文融合了 SAM，利用通道注意力机制整合特征图来选择性地强调互相关联通道的重要性，增强包含更多关键信息的特征，并抑制无关或较弱关联的特征，以平衡不同通道之间的特征信息。其

结构如图 1 所示。

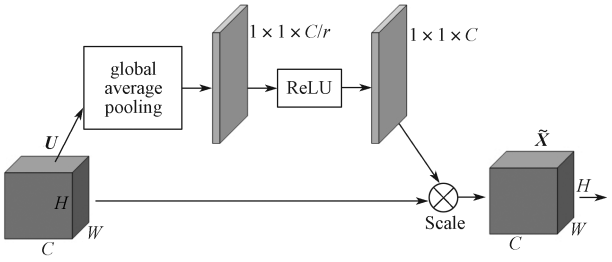


图 1　SAM 结构

Fig. 1　SAM structure

如图 1 所示，SAM 结构主要作用在残差模块的分支，将残差结构的输出作为它的输入，H、W 分别表示输入特征图的高和宽，C 表示通道数，即输入特征图大小为 $H\times W\times C$。通过全局平均池化将特征图压缩为 $1\times 1\times C$ 的向量，将全局空间信息压缩到通道描述子中，使其具有全局的感受野来对通道维度上的特征相关性进行建模。经过 1×1 卷积将特征通道数调整为输入通道数的 $1/r$，r 为缩放比例，通过实验得出 r 取 16 比较合适（详见 3.4 节），可以实现准确度和计算复杂度之间的良好平衡。对压缩了通道数的特征图经 ReLU 激活，使其具有学习通道间的非线性交互能力，再通过一个 1×1 卷积恢复通道数，最后以 Sigmoid 函数进行归一化处理，获得 0~1 之间的权重，通过 Scale 操作将每个通道赋予权重值。其中涉及的理论过程推导如式（1）所示。

$$m=F_{\mathrm{ex}}(v,W)=\sigma\left(g(v,W)\right)=\sigma\left(W_2\delta(W_1v)\right)。\quad（1）$$

式中：$F_{\mathrm{ex}}(\cdot)$ 为对经过全局平均池化后的输出特征进行激活，得到各通道特征权重值的过程；$\delta(\cdot)$ 为 ReLU 函数；W_1 和 W_2 为权重，且 $W_1\in\mathbb{R}^{\frac{C}{r}\times C}$，$W_2\in\mathbb{R}^{\frac{C}{r}\times C}$；$v$ 为特征图经过全局平均池化后的输出；$\sigma(\cdot)$ 为 Sigmoid 激活函数。

图 1 中 $\boldsymbol{U}=[u_1,u_2,\cdots,u_C]\left(\boldsymbol{U}\in\mathbb{R}^{H\times W\times C}\right)$ 在此处为 ResNet 中残差块 Residual 的输出，$\tilde{\boldsymbol{X}}=[\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_C]$ 为最终输出，是将得到的 m 映射到 U 的通道特征上，如公式（2）所示。

$$\tilde{x}_C=F_{\mathrm{Scale}}(u_C,m_C)=m_Cu_C,\qquad（2）$$

式中：$F_{\mathrm{Scale}}(u_C,m_C)$ 为标量 m_C 和特征图 $u_C\in\mathbb{R}^{H\times W}$ 的逐通道做点乘。

SAM 完成了通道相关性的构建，自适应地为不同通道学习到不同的通道注意力权重，让网络专注于有更多贡献的通道，增强判别能力。

2.2　多层次特征融合模块

在目标检测网络中，深层特征语义信息强但分辨率低，浅层特征分辨率高但语义信息弱 [21−22]，本文融合了不同分辨率的特征，利用深层特征图中含有目标丰富的语义信息和浅层特征图的局部位置信息来提高网络的性能，解决多尺度问题。

经过主干网络自底向上路径特征提取输出的 5 层特征图中，每层最后一个残差块输出的特征图为 P_1~P_5，由于 P_1 语义信息较弱、分辨率过大不利于计算，因此采用 P_2~P_5 作为加权特征融合网络的输入，用 1×1 的卷积核对原始特征横向连接，统一修正特征图的通道数为 256，进而进行自顶向下与复用的低层特征进行第一次融合，第一次只对 P_3 和 P_4 特征图做此操作，得到过渡特征集合。此时过渡特征中高层的特征信息较低层特征来说更弱，底层特征分辨率高，包含了更多小目标检测的细节信息，因此对过渡特征进行二次融合。采用 1×1 的卷积核对过渡特征图进行横向连接，对每层特征图利用下采样操作使其与更高一层特征图具有相同尺寸。再将低层特征自底向上与复用的高层特征融合，由于 P_2 特征图包含更多空间位置信息，因此将 P_2 参与二次融合，与过渡特征图 P_3 融合，P_5 特征图含有丰富的语义信息，也参与二次融合中。这种对同一层的原始输入特征直接连接到输出特征参与自底向上特征融合的做法能够充分利用高层特征的强语义信息和底层特征的空间位置信息，最终得到输出特征，送入 RPN 网络进行后续处理，MFFM 结构如图 2 所示。

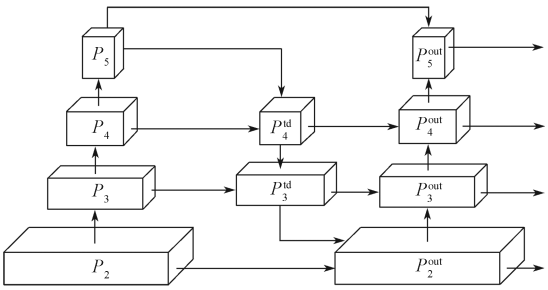


图 2　MFFM 结构

Fig. 2　MFFM structure

为了降低其复杂度，本文利用深度可分离卷积代替 3×3 普通卷积，深度可分离卷积通过两次卷积操作实现，先分别对通道同时进行 3×3 卷积，批量正则化后，通过 1×1 逐点卷积，较之普通卷积，参数量大大减少。此外，不同的输入特征具有不同的分辨率，通常它们对输出特征所作出的贡献有所不同 [23]，因此引入权重让网络学习每个输入特征的重要性，更好地平衡不同尺度的特征信息。权重计算如式（3）所示：

$$L=\sum_{i=1}^{2}\frac{w_i}{\varepsilon+\sum_{j=1}^{2}w_j}\cdot I_i,\qquad（3）$$

式中：w_i（$w_i \geqslant 0$）为可学习权重，用 ReLU 函数将权重归一化，表示每个输入特征的重要性程度；ε 为一个很小的值，设置为 0.000 1，避免数值不稳定；L 为计算的权重结果值。

以两个特征图融合为例进行说明，如式（4）（5）所示。

$$P_4^{\text{td}} = DwConv\left(\frac{w_1 \cdot P_4^{\text{in}} + w_2 \cdot Resize\left(P_5^{\text{in}}\right)}{w_1 + w_2 + \varepsilon}\right), \quad (4)$$

$$P_4^{\text{out}} = DwConv\left(\frac{w_1' \cdot P_4^{\text{td}} + w_2' \cdot Resize\left(P_3^{\text{out}}\right)}{w_1' + w_2' + \varepsilon}\right), \quad (5)$$

式（4）（5）中：P_4^{in} 为第 4 层的输入特征值；P_4^{td} 为自顶向下路径上第 4 层的过渡特征值，为便于区分同一层上不同类型的特征，本文采用上标 in、td、out 区分该层的输出特征、过渡特征、输出特征；P_4^{out} 为自底向上路径上第 4 层的输出特征；*Resize* 为用于分辨率匹配的上采样或下采样；*DwConv* 为用于特征处理的深度可分离卷积操作。

2.3 分类网络

经过特征融合的特征图通过 3 × 3 的卷积运算去除混叠效应后送到 RPN 网络。RPN 网络的详细介绍见文献 [9]，图 3 所示为 RPN 网络结构图。首先，在特征图上初步提取检测目标候选区域，本文采用 4 种不同尺度面积 {64^2，128^2，256^2，512^2}、3 种不同长宽比 {1 ∶ 2，1 ∶ 1，2 ∶ 1} 生成 12 种不同大小的 anchor，进行 1 × 1 卷积，输出 24 维的向量，输入到 Softmax 进行二分类。其次，RPN 对分类后舍弃背景的 anchor 进行边界框回归操作，得到检测目标的一系列候选区域。需要将这些候选区域映射到原图中，由于得到的候选区域大小不同，本文采用 ROI Align 方法获得固定尺寸的特征图，通过全连接和 Softmax 激活函数，并结合边界框回归进行精确地分类识别和回归定位，获得检测目标所属类别的概率和边框的准确位置。

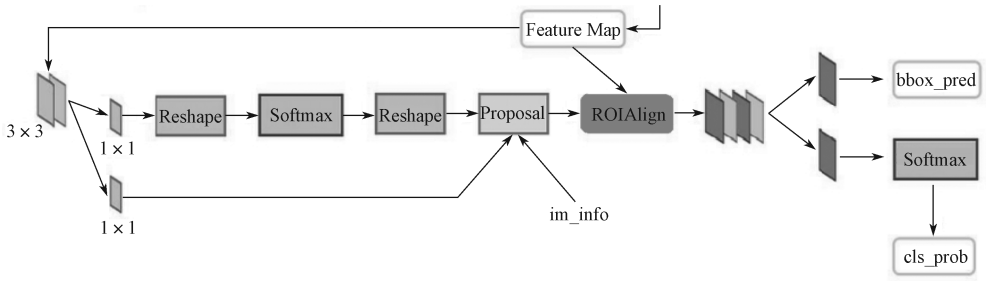


图 3 RPN 网络结构图

Fig. 3 RPN network structure

2.4 目标检测算法模型

前面设计了 SAM 和 MFFM 结构，本文基于此，提出了基于注意力机制和多层次特征融合的目标检测算法，整体框架如图 4 所示。

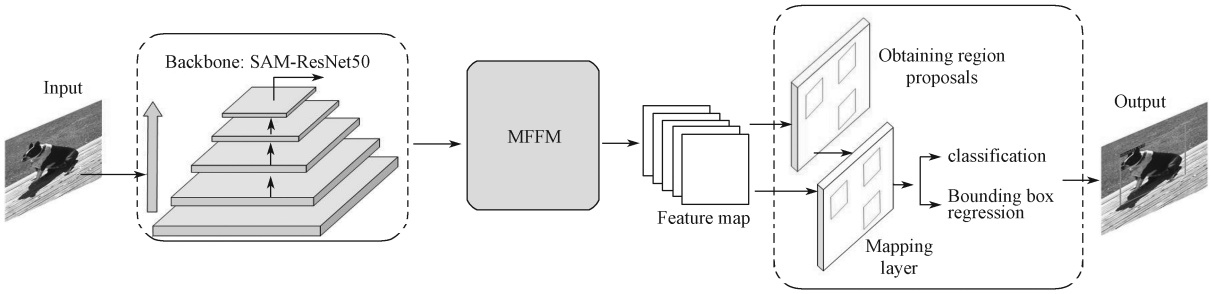


图 4 目标检测模型框架图

Fig. 4 Target detection model frame diagram

如图 4 所示，其基于 Cascade RCNN 改进，采用 ResNet50 作为主干网络，利用 SAM，旨在通过它能使网络执行动态通道特征重新校准以提高网络的表征能力，从通道维度方面提高目标检测精度。将原始图像输入主干网络完成图像特征提取，具体改进工作和实现细节详见 2.1 节。其次，本文设计了 MFFM，将主干网络提取到的不同层次特征图进行多尺度融合，实现语义特征和位置特征的有效融合，得到融合后的特征图，具体的多尺度特征融合思想和实现细节详见 2.2 节。最后利用改进的 RPN 网络 [24] 在特征图上获取包含面积种类更多的区域建议，利用 ROI Align 方法进行特征映射，用 softmax 进行分类和边框回归，其实现过程详见 2.3 节。至此完成目标检测，得到含有目标类别、目标框和置信度的图像。

3　实验和结果分析

3.1　数据集和实验环境

本实验使用 2 个数据集，一个是图像分类和目标检测中常用的标准数据集 PASCAL VOC 2012，在实验过程中将其 xml 标签文件转换为 json 格式，该数据集中共有 11 540 张已标注好的图像数据和 27 450 个目标物体，其中训练集含有 5 717 张图像数据，测试集有 5 823 张，数据集中包含行人、汽车、狗、雨伞等 20 个类别。另一个数据集是针对深度学习技术在医学图像诊断领域的应用，收集了胃肠道息肉图像数据（GP Images）。胃肠道息肉是常见的消化系统疾病，可发生于胃肠道内多个部位，会随着病情的发展出现癌变的情况，因此及时诊断发现胃肠道息肉非常重要。本文采用改进后的目标检测算法对采集到的胃肠道息肉图像进行检测和识别，协助医护人员准确捕捉到胃肠道息肉的精确位置。此数据集共有 1 000 张标注好的图像数据，包含一个类别：polyp。本实验运行环境配置：操作系统 Ubuntu 16.04，显卡 GeForce RTX 2080Ti，2.50 GHz CPU，CUDA 版本 10.2，基于 Pytorch 框架和 Python 编程语言实现。

3.2　评价指标

为了验证本文所提算法的性能，选用目标检测任务中的平均精度 *AP*（average precision）作为本文算法的评价指标，其中涉及到的精度 *p*（precision）和召回率 *r*（recall）的计算公式如式（6）所示，以检测结果框与真实框的交并比（IOU）来判定正负样本。

$$p=\frac{TP}{TP+FP},\quad r=\frac{TP}{TP+FN}。\tag{6}$$

式中：*TP* 为被模型预测为正类的正样本数；*FP* 为被预测为正类的负样本数；*FN* 为被预测为负类的正样本数。

平均精度 *AP* 则为 *PR* 曲线下的面积，计算公式如式（7）所示。

$$AP=\int_0^1 p(r)\mathrm{d}r,\tag{7}$$

式中 *p*(*r*) 为以 *r* 为参数的函数。

实验中涉及的 *AP*（平均值）、AP_{50}、AP_{75} 分别表示当 IOU 为 0.50:0.95, 0.50, 0.75 时的 *AP* 值，AP_S、AP_M、AP_L 分别表示像素面积小于 32^2, 32^2~96^2, 96^2 的目标框 *AP* 值。

3.3　数据预处理

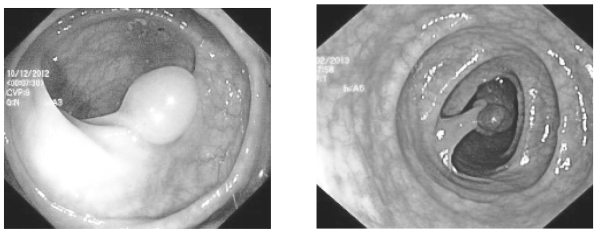
在进行目标检测时不仅要改善网络模型结构，往往还需关注数据集的质量，可对数据集进行数据增强，即对输入图像的像素点的分布、值的大小进行一些根本性的变换，同时保证图像的标签数据与之对应。本文采用在线数据增强技术中常用的随机翻转，本质上目标本身类别在翻转后未发生改变，但能增加数据集的多样性，训练过程中先对每个批次的训练数据进行数据增强，设置随机翻转概率 *flip_ratio*=0.5，即每张图像有 0.5 的概率进行翻转操作，如图 5 所示为训练集中图像数据进行翻转后的部分图像，通过对图像数据进行变换可以得到泛化能力更强的网络，能一定程度上避免网络训练过拟合的情况。

a）图像 1　b）图像 2　c）图像 3　d）图像 4　e）图像 5　f）图像 6

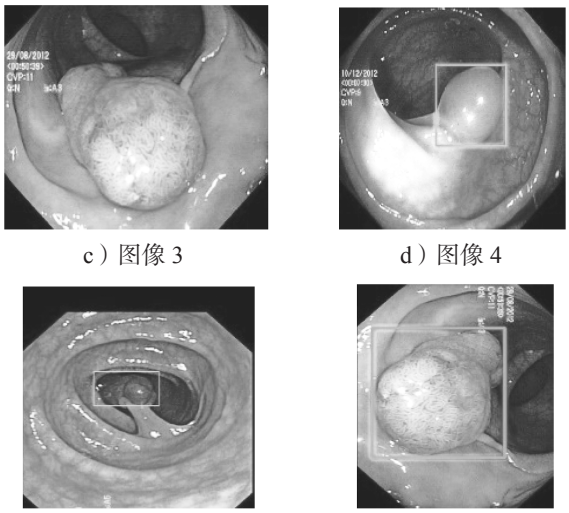
图 5　PASCAL VOC 2012 数据增强图像

Fig. 5　PASCAL VOC 2012 data image enhancement

针对胃肠道息肉数据集难以找到充足数据的问题，本文采用一些离线数据增强方法对数据集进行处理，如旋转、亮度调整、平移变换、裁剪、镜像变换。如图 6 所示为部分训练集图像原图和经过离线数据增强后的图像效果，以此增加了训练样本的数量、丰富了训练数据的分布，能够提升模型的鲁棒性。



a）图像 1　b）图像 2



c）图像 3　　d）图像 4

e）图像 5　　f）图像 6

图 6　GP Images 数据增强图像

Fig. 6　Images data image enhancement

3.4　实验结果分析

网络训练时，设置初始学习率为 0.002 5，权重衰减系数为 0.000 1，优化动量参数为 0.9。且训练时 RPN 网络中 IoU（intersection over union）阈值选为 0.7 和 0.3 来区分正负样本，测试时采用 SoftNMS 对区域建议进行分支预测，设置阈值为 0.7。为了对本文提出的检测算法的有效性进行评估，选取目标检测领域中常用的几种经典检测算法：YOLOv3、SSD、CornerNet、faster RCNN 和 Cascade RCNN，其中包含了单阶段和两阶段的方法，将其与本文提出的算法在相同的实验环境下进行训练和测试。基于所给出的评价指标，对实验结果进行对比和分析，如表 1 所示。

表 1　不同算法在 PASCAL VOC 2012 的性能比较

Table 1　Performance comparison of different algorithms on PASCAL VOC 2012

Method	Backbone	*AP*	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
YOLOv3	Darknet-53	34.3	64.0	33.2	6.9	22.2	39.4
SSD	VGG-16	37.0	64.3	37.5	4.9	18.3	44.0
CornerNet	Hourglass-104	14.8	23.1	14.6	0.6	5.8	18.7
Faster RCNN	ResNet-50	34.1	67.8	29.8	11.6	24.3	38.4
Cascade RCNN	ResNet-50	44.1	68.8	48.4	13.7	29.4	50.4
Ours	SE-ResNet-50	46.2	70.8	50.9	14.7	31.5	52.9

对比不同 IoU 阈值对应的平均精度和目标框不同像素面积对应的平均精度，可以看出，本文提出的方法明显优于其他检测算法。相较于 Cascade RCNN，分别在 *AP*、AP_{50}、AP_{75} 得到了 2.1%, 2.0%, 2.5% 的提升，对不同像素面积的目标检测精度 AP_S、AP_M、AP_L 也分别提升了 1.0%, 2.1%, 2.5%，实验证明，该方法可有效提升目标检测精度。对于数据集的 20 个类别，进一步比较了不同算法在每个类别上的平均精度 *AP*，结果如图 7 所示。

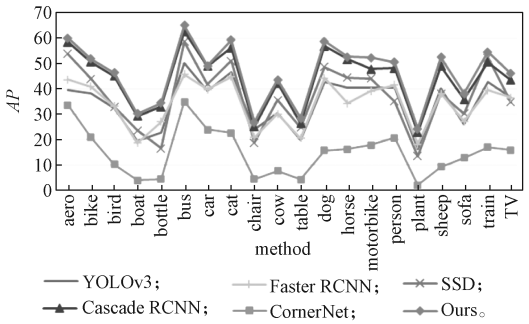


图 7　不同算法在 PASCAL VOC 2012 数据集上的检测结果

Fig. 7　Detection results of different algorithms on Pascal VOC 2012 dataset

图 7 中数据表明，相较于其他目标检测算法，本文提出的基于注意力机制和多层次特征融合的检测方法，在精度上有显著提升，该方法在每个类别上的平均精度都优于其他方法，证明了该方法的有效性。

除了和目标检测方法中经典算法相比之外，本文还选取了近年来采用其它注意力机制和特征融合方式的算法，在 PASCAL VOC 2012 数据集上训练和测试，并将测试结果与本文提出的算法进行比较，结果见表 2。表中 PANet、HRNet、NAS-FPN 采用了其它多尺度特征融合方法，DANet、ACNet 中则是引入了其它注意力机制方法。由表可知本文提出算法与上述算法相比，检测精度都有所提升。其中 HRNet、ACNet 在 AP_S 上的检测结果分别为 14.9%, 14.8%，虽然本文提出的算法在 AP_S 上略低，但在 *AP*、AP_{50}、AP_{75}、AP_M、AP_L 上的检测结果远优于其它算法，总体上本文的算法能够实现较好的检测效果。

表 2　相似算法在 PASCAL VOC 2012 上的性能比较

Table 2　Performance comparison of similarity algorithms on PASCAL VOC 2012

Method	*AP*	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
PANet	44.8	70.2	49.2	13.7	30.5	51.0
HRNet	44.3	68.5	49.3	14.9	30.3	50.5
NAS-FPN	40.9	61.2	44.5	11.6	24.9	47.1
DANet	44.0	68.4	48.9	14.1	30.9	50.0
ACNet	45.1	69.5	50.1	14.8	31.0	51.9
Ours	46.2	70.8	50.9	14.7	31.5	52.9

为了验证本文所提出的方法中各个模块对检测性能的优化作用，分别对 SAM 和 MFFM 的有效性进行评估，并基于相同的实验环境和参数配置，在 PASCAL VOC 2012 数据集上进行消融实验，分析实验结果。具体实验方案如下：1）在模型中单独验证 SAM；2）在模型中单独验证 MFFM；3）在模型中同时验证 SAM 和 MFFM。

消融实验的数据结果如表 3 所示，本文提出的方法是在 Cascade RCNN 模型的基础上进行改进，

因此相较于表 1 中未添加 SAM 和 MFFM 的 Cascade RCNN 的检测结果，单独添加通道注意力模块后的模型在 AP_{50}、AP_{75} 上分别得到 0.7%, 0.8% 的提升，单独添加了多层次特征融合方法的模型在 AP、AP_{50}、AP_{75} 上分别得到 0.2%, 1.0%, 0.8% 的提升，当 SAM 和 WFFM 都运用到模型中时，精度的提升效果最明显，实验结果表明使用 SAM 和 MFFM 能够有效提升模型检测精度。

表 3　在 PASCAL VOC 2012 数据集上的消融实验结果

Table 3　Ablation experimental results on PASCAL VOC 2012 dataset

Method	+SAM	+MFFM	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Our	√		43.6	69.2	47.7	14.3	29.6	49.8
		√	44.3	69.5	49.2	13.5	30.1	50.7
	√	√	46.2	70.8	50.9	14.7	31.5	52.9

对于 GP Images 数据集经过同样的策略训练和测试，表 4 给出了不同像素面积的目标检测精度。表 4 中实验数据表明，相较于 Cascade RCNN ，本文提出的算法在 AP、AP_{50}、AP_{75} 上分别提升了 5.8%, 1.5%, 1.1%，能够实现较好的检测效果。

表 4　不同算法在 GP Images 上的性能比较

Table 4　Performance comparison of different algorithms on GP Images

Method	Backbone	AP	AP_{50}	AP_{75}
YOLOv3	Darknet-53	52.5	83.3	61.8
SSD	VGG-16	60.0	84.3	67.0
CornerNet	Hourglass-104	50.4	64.4	51.9
Faster RCNN	ResNet-50	51.1	83.2	57.1
Cascade RCNN	ResNet-50	55.9	84.7	66.0
Ours	SE-ResNet-50	61.7	86.2	67.1

为了验证对 GP Images 数据集采取数据增强策略的作用，本文将所提出的模型在原 GP Images 数据集和采取了数据增强策略的 GP Images 数据集上分别进行训练和测试，以不同像素面积的检测精度作为评价指标，实验结果如表 5 所示。由表中数据可知，采取了数据增强策略得到的检测效果有一定的提升，在 AP 和 AP_{75} 上分别提升了 3.4% 和 0.8%，对网络模型的检测精度具有优化作用。

表 5　数据增强策略在 GP Images 上的性能对比

Table 5　Performance comparison of data enhancement strategies on GP Images

Method	Data Enhancement	AP	AP_{50}	AP_{75}
Cascade RCNN	无	55.9	84.7	66.0
	有	59.3	84.1	66.8

最后，对 2.1 节简单的注意力模块（SAM）中涉及的缩放比例 r 的取值进行实验验证，本实验在 Cascade RCNN 模型中添加 SAM，且 r 的取值分别为 4, 8, 16, 32，在 GP Images 数据集上经过训练测试，实验结果如表 6 所示。由表可知，当 r=16 时，模型在精度和速度方面较好平衡，故本文 r 取 16。

表 6　不同缩放比例 r 对模型性能的影响

Table 6　Effects of different scaling ratios r on model performance

Method	r	AP	AP_{50}	AP_{75}	Speed/FPS
Cascade RCNN with SAM	4	57.3	81.6	62.2	76
	8	56.4	81.3	61.9	77
	16	59.8	83.5	64.1	77
	32	59.8	81.5	64.6	76

4　结语

本文提出了基于注意力机制和多层次特征融合的目标检测算法，通过在主干网络中融合 SAM，利用通道注意力机制有选择的突出作用性更强的通道特征信息，从而提高网络的判别能力；其次，本文针对目标检测算法中的多尺度问题，改进了 FPN，结合深度可分离卷积，设计了 MFFM，充分融合深层特征丰富的全局语义信息和浅层特征的局部空间位置信息，使网络提取的特征更具表征能力，并为不同层次的特征引入权重，更好地平衡不同尺度的特征信息。实验结果表明，本文提出的算法在一定程度上大大提高了目标检测精度，改善了检测效果。接下来将进一步优化模型，致力于在保持精度的同时提升网络的效率。

参考文献：

[1] 董小伟，韩　悦，张　正，等 . 基于多尺度加权特征融合网络的地铁行人目标检测算法 [J]. 电子与信息学报，2021，43(7)：2113−2120.
DONG Xiaowei，HAN Yue，ZHANG Zheng，et al. Metro Pedestrian Detection Algorithm Based on Multi-Scale Weighted Feature Fusion Network[J]. Journal of Electronics & Information Technology，2021，43(7)：2113−2120.

[2] 张瑞倩，邵振峰，PORTNOV A，等 . 多尺度空洞卷积的无人机影像目标检测方法 [J]. 武汉大学学报·信息科学版，2020，45(6)：895−903.
ZHANG Ruiqian，SHAO Zhenfeng，PORTNOV A，et al. Multi-Scale Dilated Convolutional Neural Network for Object Detection in UAV Images[J]. Geomatics and Information Science of Wuhan University，2020，45(6)：895−903.

[3] 谭台哲，卢剑彪，温捷文，等 . 应用卷积神经网络与 RPN 的交通标志识别 [J]. 计算机工程与应用，2018，54(21)：251−256，264.
TAN Taizhe，LU Jianbiao，WEN Jiewen，et al. Traffic Signs Recognition Applying with Convolutional

Neural Network and RPN[J]. Computer Engineering and Applications，2018，54(21)：251−256，264.

[4] DENG J，XUAN X J，WANG W F，et al. A Review of Research on Object Detection Based on Deep Learning[J]. Journal of Physics：Conference Series，2020，1684(1)：012028.

[5] PAPAGEORGIOU C P，OREN M，POGGIO T. A General Framework for Object Detection[C]//Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271). Bombay：IEEE，1998：555−562.

[6] DALAL N，TRIGGS B. Histograms of Oriented Gradients for Human Detection[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego，CA，USA：IEEE，2005：886−893.

[7] LOWE D G. Distinctive Image Features from Scale-Invariant Keypoints[J]. International Journal of Computer Vision，2004，60(2)：91−110.

[8] GIRSHICK R，DONAHUE J，DARRELL T，et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[J]. 2014 IEEE Conference on Computer Vision and Pattern Recognition，2014：580−587.

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

[10] 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. Salt Lake City：IEEE，2018：6154−6162.

[11] LIU W，ANGUELOV D，ERHAN D，et al. SSD：Single Shot Multibox Detector[C]//2016 European Conference on Computer Vision. Amsterdam：Springer，2016：21−37.

[12] REDMON J，DIVVALA S，GIRSHICK R，et al. You only Look Once：Unified，Real-Time Object Detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas：IEEE，2016：779−788.

[13] 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. Honolulu：IEEE，2017：936−944.

[14] GHIASI G，LIN T Y，LE Q V. NAS-FPN：Learning Scalable Feature Pyramid Architecture for Object Detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach：IEEE，2019：7029−7038.

[15] LIU S，QI L，QIN H F，et al. Path Aggregation Network for Instance Segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City：IEEE，2018：8759−8768.

[16] SUN K，XIAO B，LIU D，et al. Deep High-Resolution Representation Learning for Human Pose Estimation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach：IEEE，2019：5686−5696.

[17] TAN M X，PANG R M，LE Q V. EfficientDet：Scalable and Efficient Object Detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle：IEEE，2020：10778−10787.

[18] CAO J X，CHEN Q，GUO J，et al. Attention-Guided Context Feature Pyramid Network for Object Detection[EB/OL]. [2022−02−23]. https://arxiv.org/abs/2005.11475.

[19] XING H J，WANG S，ZHENG D Z，et al. Dual Attention Based Feature Pyramid Network[J]. China Communications，2020，17(8)：242−252.

[20] HU J，SHEN L，SUN G. Squeeze-and-Excitation Networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City：IEEE，2018：7132−7141.

[21] 王凤随，陈金刚，王启胜，等 . 自适应上下文特征的多尺度目标检测算法 [J]. 智能系统学报，2022，17(2)：276−285.
WANG Fengsui，CHEN Jingang，WANG Qisheng，et al. Multi-Scale Target Detection Algorithm Based on Adaptive Context Features[J]. CAAI Transactions on Intelligent Systems，2022，17(2)：276−285.

[22] ZHANG Y，CHEN Y M，HUANG C，et al. Object Detection Network Based on Feature Fusion and Attention Mechanism[J]. Future Internet，2019，11(1)：9.

[23] 张世辉，王红蕾，陈宇翔，等 . 基于深度学习利用特征图加权融合的目标检测方法 [J]. 计量学报，2020，41(11)：1344−1351.
ZHANG Shihui，WANG Honglei，CHEN Yuxiang，et al. An Object Detection Method Based on Deep Learning Using Feature Map Weighted Fusion[J]. Acta Metrologica Sinica，2020，41(11)：1344−1351.

[24] 黄友文，冯 恒，万超伦 . 基于区域生成网络结构的多层特征融合目标检测算法 [J]. 科学技术与工程，2019，19(24)：213−217.
HUANG Youwen，FENG Heng，WAN Chaolun. Multi-Feature Fusion Object Detection Algorithm Based on Region Proposal Network Structure[J]. Science Technology and Engineering，2019，19(24)：213−217.

（责任编辑：申 剑）