文章编号：1005-5630(2023)06-0014-11 **DOI**: 10.3969/j.issn.1005-5630.202302030011

MSA-Net:一种基于多阶段注意力机制的少样本目标检测方法

汤应薇，张荣福，丁 然，张 杰
（上海理工大学 光电信息与计算机工程学院，上海 200093）

摘要：近年来，样本较少场景下的目标检测引起了广泛的关注。由于少样本提供的信息有限，大部分少样本目标检测模型采用改进的 Faster RCNN 检测框架进行研究。但由于 Faster RCNN 框架中潜在的模块矛盾问题，现有的少样本目标检测模型的特征捕捉和分类的能力有待提高。为解决以上问题，以 Faster RCNN 框架为基础，加入了梯度反传解耦机制，缓解在反向传播过程中，RPN 和 RCNN 的冲突对主干网络的负面影响。为提高目标检测模型的特征捕捉能力，采用元学习框架，并融合基于注意力机制的蒸馏模块和多尺度注意力模块，充分利用查询集和支持集的信息，捕捉更多全局特征信息。大量的实验证明，在随机采样目标数 k=1, 2, 3, 5, 10 设置下，改进后的模型在 Pascal VOC 数据集的新类上，分别达到 21.8%，34.7%，40.9%，44.5%，51.7% mAP（AP50）。在 k = 10, 30 设置下，改进后的模型在 COCO 数据集的新类上，分别达到 25.1%，27.6% mAP(AP50)。

关键词：深度学习；少样本学习；目标检测
中图分类号：TP 312 **文献标志码**：A

MSA-Net: few-shot object detection with multi-stage attention mechanism

TANG Yingwei，ZHANG Rongfu，DING Ran，ZHANG Jie
（School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China）

Abstract: In recent years, object detection in scenarios with fewer samples has attracted widespread attention. Due to the limited information provided by the few samples, most of few-shot object detection models are studied using the improved Faster RCNN detection framework. However, due to the potential module contradiction problem in the Faster RCNN framework, the feature capture and classification capabilities of the existing few-shot object detection models need to be improved. In order to solve the above problems, this paper adds a gradient decoupling

收稿日期：2023-02-03
基金项目：基础科研条件与重大科学仪器设备研发计划（2022YFF0706003）
第一作者：汤应薇 (1998—)，女，硕士研究生，研究方向为计算机视觉、少样本学习。E-mail：1539861596@qq.com
通信作者：张荣福 (1971—)，男，教授，研究方向为智能检测技术。E-mail：zrf@usst.edu.cn

mechanism based on the Faster RCNN framework to alleviate the negative impact of the conflict between RPN and RCNN on the backbone network during the backpropagation process. In order to improve the feature detection ability of the object detection model, this paper adopts meta-learning framework, integrates the distillation module based on attention mechanism and the multi-scale attention module, and makes full use of the information of the query set and support set to capture more global feature information. A large number of experiments have proved that under the setting of randomly sampled shot amount k=1, 2, 3, 5, 10, the improved model can reach 21.8%, 34.7%, 40.9%, 44.5%, 51.7% mAP (AP50) on the new class of Pascal VOC dataset, respectively. Under the k=10, 30 setting, the improved model achieves 25.1% and 27.6% mAP (AP50) on the new class of the COCO dataset, respectively.

Keywords: deep learning；few-shot learning；object detection

引　言

随着计算机计算性能的提高和深度学习技术的发展，基于传统图像算法进行目标检测的方法逐渐被取代，利用深度卷积网络的框架开始成为主流。但目前深度学习框架过度依赖大量有标注的数据集和迭代次数、学习率等超参数的调节。在现实中，也受各种客观条件的限制，获取大量标签[1]的实验样本数据需要耗费大量的人力物力[2]。鉴于以上难题，研究模型通过少量样本数据训练，能实现快速收敛并在大量样本数据中具有高度泛化性能显得尤为重要。根据对人类认知模式的观察与研究，人类可以通过极少量的图像样本来准确识别一个新物体的所属类别。受这种快速学习方式[3]的启发，基于少样本学习概念的图像处理方法被提出。但目前大部分少样本问题的相关工作集中在图像分类问题上[4–7]，对于少样本目标检测问题的研究工作相对较少。由于目标检测问题不仅包括对图像类别的预测，还涉及到对目标的定位问题，这使该任务与少样本分类任务相比复杂度更高。

目前少样本目标检测任务主要采用少样本学习方法结合成熟的深度学习目标检测框架。在少样本学习方法研究初期，大多数模型采用两阶段微调方法提升少量新样本的检测精度。Chen等[8]采用多层深度卷积设计边界框回归策略，抑制在基类训练过程中的过度拟合现象。Wang等[9]在少样本微调阶段，固定特征提取层权重参数，仅微调检测模型的最后一层参数，提高少样本的检测精度和准确率（5%～10%）。但由于在基类训练过程中过度拟合，两阶段微调的方法在微调阶段对新类的检测精度欠佳。因此，后续的少样本目标检测研究融入元学习框架[9–11]，将元学习器加入现有的目标检测网络，提高后续特征提取层的有效性。但由于少样本目标检测任务的挑战性，文献 [9–11] 采用的元学习方式，只是对支持集特征进行全局池化，丢失了局部特征的详细上下文信息，导致无法学习到预测分类和定位的关键特征。因此，后续的研究将两阶段微调和元学习方式结合，提出了 Meta-SSD 模型[12]。该模型通过一系列的少样本目标检测任务学习得到一个元学习器，来指导检测器在新任务中快速且准确地更新参数。Meta RCNN 模型[11]将元学习方法和 Faster RCNN 框架结合，利用元学习器学习到的可共享参数，帮助新类检测器高效提取特征。

但 Faster RCNN 作为一种两阶段检测体系，类无关的区域建议网络（RPN）和类相关的区域卷积神经网络（RCNN）通过共享主干交换优化信息时，可能会因为优化目标的不同产生冲突，导致检测能力下降。文献 [13] 提出，分类网络注重类间差异的特征，定位网络注重类内差异的特征。不匹配的特征可能会影响 RPN 产生许多低质量的目标分数，导致分类能力下降。文献 [14–16] 研究了特征尺度变化对模型检测精度的影响，由于锚匹配机制，经过特征金字塔后负样本会增加。少量的标注样本也会导致特征提取网络只能提取单一的特征尺度，使得模型对多尺度图像的检测性能下降。在基类训练过程中，丰富

的数据可以降低负样本产生的负面影响，但在元微调过程中，新类样本量小，负样本会导致模型对新类学习能力下降。

为解决以上问题，本文基于 Faster RCNN 架构，提出了 MSA-Net 模型，这一模型对 Faster RCNN 进行了 3 个方面的改进。(1)加入梯度反传解耦机制：在反向传播过程中，同时对 RPN 和 RCNN 进行梯度解耦，缓解在反向传播过程中 RPN 和 RCNN 的冲突对主干网络的负面影响。(2)加入注意力蒸馏模块，利用键–值映射方式，提取支持集和查询集之间的关联信息，过滤掉大多数背景框或类别无关特征。(3)在 ROI 池化阶段提出多尺度注意力模块，提取 3 种不同尺寸的细粒度特征，并对特征采用自主意力机制，使模型对不同尺度的特征具备更好的泛化能力，便于后续的分类器和回归器工作。大量的实验证明，在 $k=1, 2, 3, 5, 10$ 的设置下，改进后的模型在 Pascal VOC[17] 的新类上，分别达到 21.8%，34.7%，40.9%，44.5%，51.7% 平均精度(mAP)(AP50)。在 $k=10, 30$ 设置下，改进后的模型在 COCO[18] 数据集的新类上，分别达到 25.1%，27.6% mAP(AP50)。

1　相关定义

如图 1 所示，遵循文献 [11, 19, 20] 中的训练范式，基于元学习的少样本目标检测训练包括元训练和元微调阶段。结合文献 [21] 对元特征学习器 $\mathcal{D}$ 的设计，对于一张输入的查询图像 q，通过抽样构建对应的支持集 S，支持集 S 包括所有类别的图像–掩膜对，$S=\{I_i, M_i\}_{i=1}^{N}$，其中 $I_i \in \mathbb{R}^{h\times w\times 3}$ 是 RGB 图像，$M_i \in \mathbb{R}^{h\times w}$ 是根据标注边界框生成的掩码图像，N 表示类的数量。具体地说，在元训练阶段，先利用有规范标注的基类数据 D_{base} 通过元特征学习器生成查询集–支持集图像对，来训练设计的模型。在元微调阶段，加入样本量少的新类并随机抽选 K 个样本形成新数据集 D_{novel}，再采用相同的元特征学习器方式构

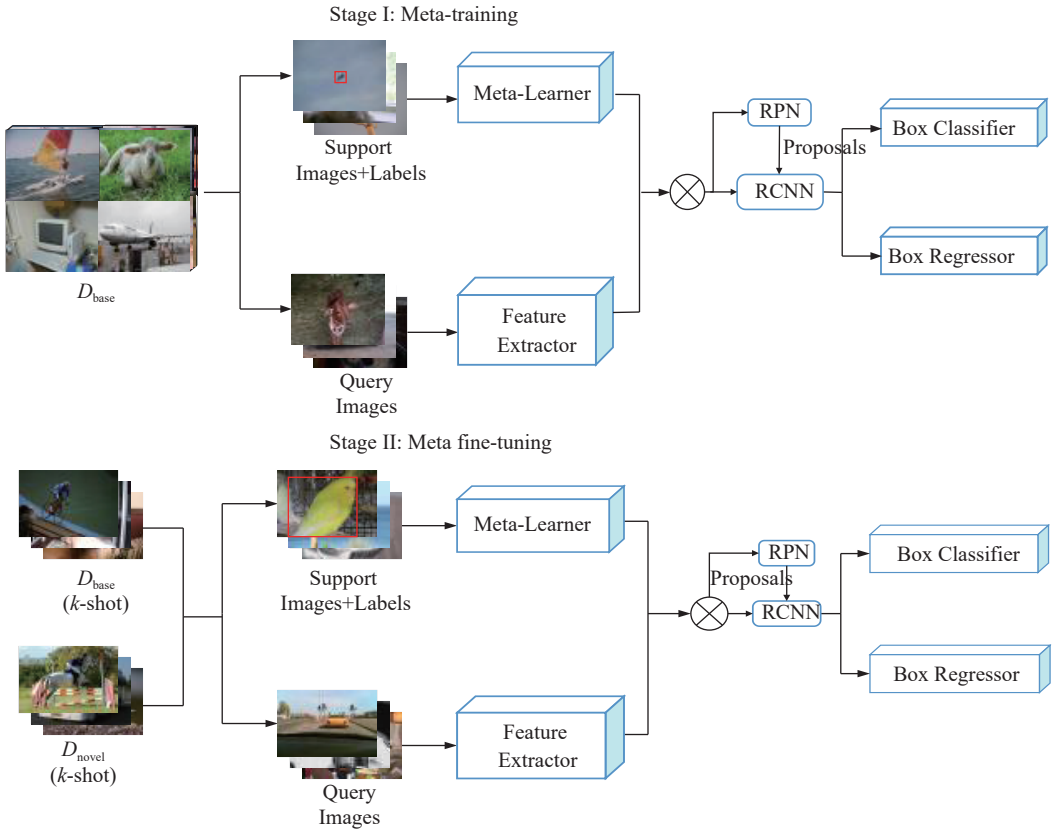

图 1　基于元学习的少样本目标检测框架的学习策略

Fig. 1　Learning strategy for a few-shot object detection framework based on meta-learning

建训练数据集。元微调训练过程与元训练阶段相同，但微调过程模型收敛速度快，训练次数相对很少。

总体来说，整体模型的输入是一张查询集图像 q 和从训练集随机抽样得到的支持对，输出是对查询集图像的检测预测。与文献 [20] 不同，本文并不采用特征重加权模块，而是借鉴文献 [11] 的方法，采用共享主干网络参数的形式对支持对进行特征提取。

2　MSA-Net

图 2 介绍了 MSA-Net 模型的体系结构。首先，数据被分为支持集和查询集两部分输入特征提取层，特征提取层采用多尺度特征金字塔结构和 ResNet-101 网络。MSA-Net 基于 Faster RCNN 结构框架，加入梯度反传解耦机制、注意力蒸馏模块以及多尺度注意力模块。

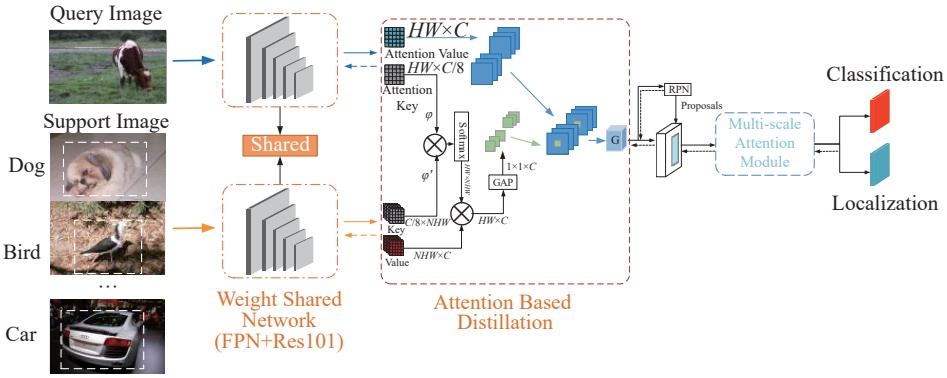

图 2　MSA 模型的总体框架

Fig. 2　General framework of the MSA model

2.1　梯度反传解耦机制

Faster RCNN 作为一种两阶段检测结构，根据功能可分为 3 个组成部分，即提取广义特征的主干网络，用于生成类无关建议的高效 RPN，和执行类相关分类和定位任务的 RCNN 头。输入图像经过主干网络生成特征映射后，被并行输入 RPN 和 RCNN。如图 3 的梯度流所示，在传统的 Faster RCNN 中，RPN 与 RCNN 在反向传播过程中，通过共享的主干网络的梯度流，交换优化信息。但 RPN 在类不可知的前提下，优化目标是产生定位准确的目标框。RCNN 的优化目标是在已知类相关的前提下，进行目标的定位。这与 RPN 的优化目标存在潜在的矛盾，会导致在少样本情况下，模型检测能力下降。受梯度反传过程的启发，本文设计在梯度反传过程中对 RPN 和 RCNN模块进行解耦（Decoupled-layer），如图 4 所示。首先在前向传播过程中，对两个模块分别添加线性变换层 A，该线性变换层由权重 ω 和偏移 b 两个可学习变量组成，目的是通过可学习参数的调节，增强特征表示并进行前向解耦。

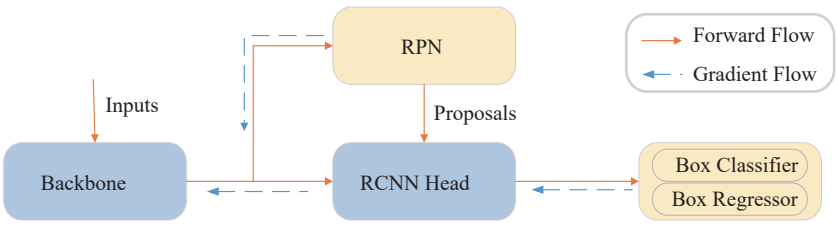

图 3　传统的 Faster RCNN 架构[22]

Fig. 3　Traditional Faster RCNN architecture [22]

在梯度反传过程中，为调节 3 个模块之间的解耦程度，将 RPN 和 RCNN 反传的梯度 $\mathbb{G}_{\text{rpn}}$ 和 $\mathbb{G}_{\text{rcnn}}$，分别乘以解耦系数 λ_{rpn}, λ_{rcnn} 向前一层传播。

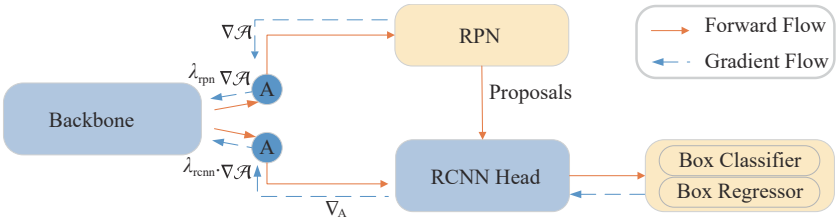

图 4　梯度反传解耦机制

Fig. 4　Gradient backpropagation decoupling mechanism

$$\mathcal{A}(\mathrm{x}) = \omega x + b \tag{1}$$

$$\frac{\mathrm{d}\mathbb{G}_{\mathrm{rpn}}}{\mathrm{d}x} = \lambda_{\mathrm{rpn}}\nabla\mathcal{A} \tag{2}$$

$$\frac{\mathrm{d}\mathbb{G}_{\mathrm{rcnn}}}{\mathrm{d}x} = \lambda_{\mathrm{rcnn}}\nabla\mathcal{A} \tag{3}$$

因此，模型的总优化目标为

$$\begin{aligned}&\arg\min_{\Theta}\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{\mathrm{RPN}}(F_{\mathrm{RPN}}(\mathbb{G}_{\mathrm{rpn}}(F_{\mathrm{b}}(x;\theta_{\mathrm{b}}));\theta_{\mathrm{RPN}}),\\&y_{\mathrm{RPN}})+\eta\mathcal{L}_{\mathrm{RCNN}}(F_{\mathrm{RCNN}}(\mathbb{G}_{\mathrm{rcnn}}(F_{\mathrm{b}}(x;\theta_{\mathrm{b}}));\theta_{\mathrm{RCNN}}),\\&y_{\mathrm{RCNN}}),\ \Theta=\{\theta_{\mathrm{b}},\theta_{\mathrm{RPN}},\theta_{\mathrm{RCNN}}\}\end{aligned} \tag{4}$$

式中：N 是训练样本的数量；$\mathcal{L}_{\mathrm{RPN}}$ 和 $\mathcal{L}_{\mathrm{RCNN}}$ 是 RPN 和 RCNN 的网络损失函数；$\theta_{\mathrm{b}},\theta_{\mathrm{RPN}}$ 和 θ_{RCNN} 分别是主干网络、RPN 和 RCNN 的学习参数；η 为平衡 $\mathcal{L}_{\mathrm{RPN}}$ 和 $\mathcal{L}_{\mathrm{RCNN}}$ 的超参数，一般设置为 1。具体地说，添加的解耦机制并不会影响到 RPN 和 RCNN 的优化，但主干网络的梯度下降会受机制的影响，可描述为

$$\theta_{\mathrm{b}} \leftarrow \theta_{\mathrm{b}} - \gamma\left(\lambda_{\mathrm{rpn}}\frac{\partial L_{\mathrm{rpn}}}{\partial\theta_{\mathrm{b}}}+\lambda_{\mathrm{rcnn}}\frac{\partial L_{\mathrm{rcnn}}}{\partial\theta_{\mathrm{b}}}\right) \tag{5}$$

式中：γ 表示学习率；λ_{rpn} 和 λ_{rcnn} 的取值对主干网络参数 θ_{b} 的更新有深刻影响。

2.2　注意力蒸馏模块

先前的工作[12, 20]采用元学习的方式，对支持集特征进行平均池化操作，抽取类向量来指导查询集元特征学习。但即使是同一类别的样本，输入的查询集和对应的支持集样本，也存在巨大差异（如：尺寸、外观、位置等）。这会对后续的特征学习产生负面影响。面对以上这些问题，本文采用了基于注意力机制的蒸馏模块，从支持集特征中尽量提取足够的细节并通过支持集信息过滤掉无关的背景和负面信息，帮助后续的 RPN 和 RCNN 进行预测。

键–值对嵌入：在查询集和对应的支持集图像通过共享参数的特征提取器提取特征后，支持集和查询集特征进入注意力蒸馏模块。首先，支持集和查询集样本通过特征提取层提取特征，再分别通过设计的深度编码器生成各自的键–值对。深度编码器由两个并行的 3×3 卷积层组成，对输入特征分别生成键和值的特征映射。虽然两部分的深度编码器采用相同的参数结构，但并不共享参数。键映射应用于检测支持集和查询集特征之间的相似性，而值映射储存了特征的全部信息。通过键的编码和得到的相似性权重，帮助查询集特征和支持集值映射进行匹配和识别。因此，对于查询特征，输出为键–值对特征映射：$k_{\mathrm{q}}\in\mathbb{R}^{\frac{C}{8}\times H\times W}$，$v_{\mathrm{q}}\in\mathbb{R}^{C\times H\times W}$，其中，$C$ 为特征的维数，W 和 H 分别为特征的长和高。对于支持集特征，每个类的特征都独立生成键–值对映射，输出是：$k_{\mathrm{s}}\in\mathbb{R}^{N\times\frac{C}{8}\times H\times W}$，$v_{\mathrm{s}}\in\mathbb{R}^{N\times C\times H\times W}$，其中 N 是支持集中类的数量。随后，生成的键–值对被送入关系蒸馏部分，在这一部分中，支持集和查询集的键–值对将被密集匹配，以找到目标对象。

键–值关系蒸馏：在获取映射后，执行键–值关系蒸馏。如图 5 所示，通过计算支持集和查询集键-值的相似度来评估支持集特征的软权重，对特征映射的每一像素执行相似度计算，公式为

$$S\left(\boldsymbol{k}_{\mathrm{q}i},\boldsymbol{k}_{\mathrm{s}j}\right)=\varphi\left(\boldsymbol{k}_{\mathrm{q}i}\right)^{\mathrm{T}}\varphi'\left(\boldsymbol{k}_{sj}\right) \tag{6}$$

式中：i 和 j 表示查询集和支持集特征的索引号；φ，φ' 表示查询集和支持集因训练过程不断变化的参数而形成的动态学习线性相似函数。对像素级特征相似度通过 softmax 函数进行归一化来计算匹配权重 $\mathcal{W}$。

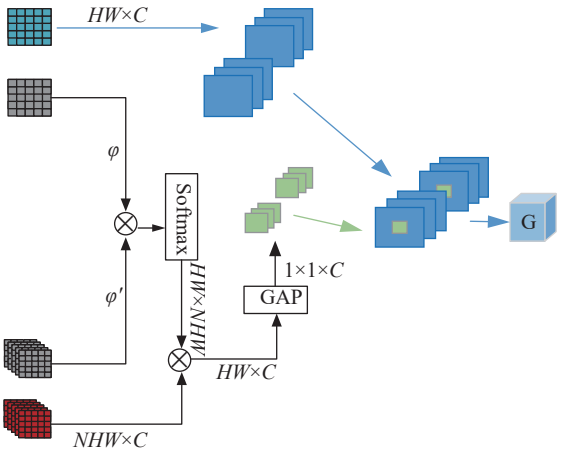

图 5　注意力蒸馏模块

Fig. 5　Attention based distillation module

$$\mathcal{W}_{ij}=\frac{\exp\left(S\left(\boldsymbol{k}_{\mathrm{q}i},\boldsymbol{k}_{\mathrm{s}j}\right)\right)}{\sum_{j}\exp\left(S\left(\boldsymbol{k}_{\mathrm{q}i},\boldsymbol{k}_{\mathrm{s}j}\right)\right)} \tag{7}$$

然后，对支持集特征通过生成的软权重进行加权。

$$x=\mathcal{W}*v_{\mathrm{s}} \tag{8}$$

式中 $*$ 表示矩阵内积。值得注意的是，支持集包含 N 个特征，因此产生了 N 个键–值对，在这里对 N 个输出的结果进行求和获得最终结果。通过平均池化层（GAP），$x\in\mathbb{R}^{H\times W\times C}$ 变为 $1\times1\times C$ 的向量。最终，查询集特征 $y\in\mathbb{R}^{H\times W\times C}$ 与支持集池化后的特征 x 通过深度交叉相关（Depth-wise Cross Correlation）模块生成最终的注意力特征图。因此，最终的输出公式为

$$G_{H,W,C}=\sum_{i,j}x_{i,j,C}\dot{y}_{H+i-1,W+j-1,C},\quad i,j\in\{1,\cdots,S\} \tag{9}$$

式中 G 表示最终的注意力特征图。对于支持集最终得到的特征 x，被用作内核以深度交叉相关方式在查询集特征 y 上滑动。

2.3　多尺度注意力机制模块

RCNN 以 RPN 的区域建议和特征作为输入，经过 ROI 对齐后进行特征池化，用于最后的分类和边界框回归。大部分研究[9–12, 20]使用固定的分辨率 8 来进行特征池化操作，但这一方式会导致部分关键信息的丢失。传统的目标检测方法往往有大量数据进行基础训练，这种方式可以弥补关键信息的丢失。但少样本条件下，由于每一类的样本只有几张图像，这导致尺度变化对模型检测效果有很大的影响。模型因为泛化能力不够，无法适应新类不同尺度的特征。因此，本文提出多尺度注意力模块（图 6），选择 4，8，12 三种分辨率能帮助模型对不同尺寸的特征具有更好的泛化能力。较大的分辨率通过放大特征，可以捕捉特征详细的上下文语义信息，便于对小目标的识别和检测，而较小的分辨率通过缩放特征尺寸，能捕捉特征的全局信息，有利于识别较大的目标。

除此以外，添加了 1 种自注意力机制（SE-

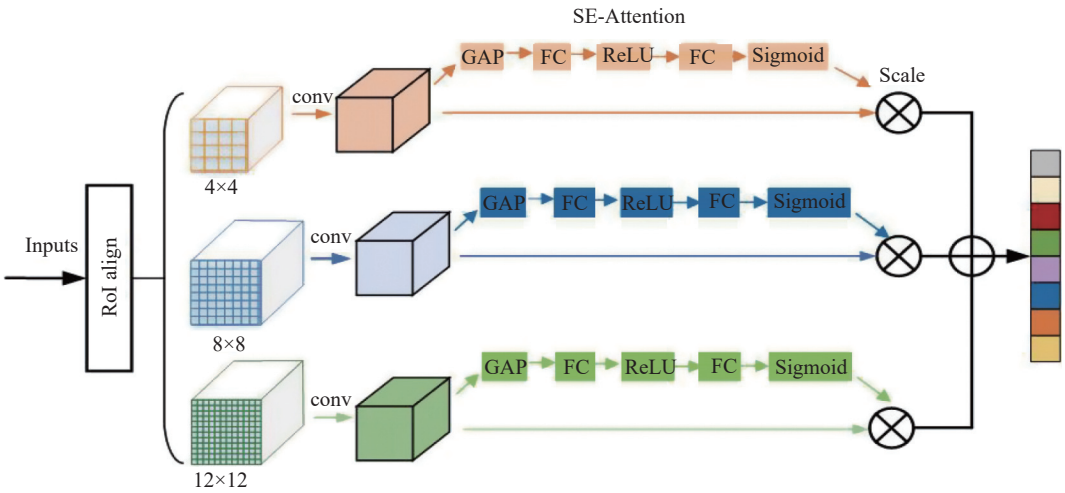

图 6　多尺度注意力模块

Fig. 6　Multi-scale attention module

Attetion)[23] 对提取的不同分辨率的特征进行特征增强，再进行有效的聚合。SE-Attention 为每个尺度的特征建立注意力分支。对于每一分支，首先进行冻结操作，利用全局池化层获得全局的感受野，然后利用两个全连接层和 ReLU 函数进行激活操作，为每个特征通道生成权重，再利用 Sigmoid 函数获得 0～1 之间的归一化权重，进行重加权，将通道权重加权到先前的特征上。最终，这一模块的输出是不同尺寸特征的加权和。

3　实　验

在本节中，对提出的 MSA-Net 模型进行了全面的实验评估来检测模型在少样本目标检测方面的作用。3.1 节介绍了数据集的详细设计和训练策略。3.2 节对 MSA-Net 模型在 Pascal VOC 数据集和 COCO 数据集上的表现进行了详细的实验分析。3.3 节对提出的创新模块进行消融实验结果分析。

3.1　数据集和学习策略介绍

数据集：为与其他先进的方法进行公平比较，根据文献 [9] 的方式构建少样本数据集。对于 Pascal VOC 数据集，将 VOC 2007trainval set 和 VOC 2012trainval set 作为训练集，VOC2007test set 作为测试集。评估指标为平均精度（mAP）。具体地，数据集根据目标类型将 20 个类随机分为 15 个基类和 5 个新类。首先在元训练阶段，对基类所有数据进行训练，再在元微调阶段，对每一类别从训练集中随机采样 k（k = 1, 2, 3, 5, 10）个目标。为保证结果的公平性，对类别进行了 3 次随机分组。对于 COCO 数据集，将与 Pascal VOC 数据集不相交的 60 个类作为基类，剩余的 20 个类作为新类。使用验证集中标记为 minival 的数据作为测试集，其余的验证集和训练集数据作为模型的训练集。训练阶段的设置和 Pascal VOC 数据集训练阶段相同，k 设置为 10, 30。

学习策略：首先，使用 Faster RCNN 框架[22] 作为基本网络架构，采用具有特征金字塔网络[14] 的 ResNet-101 作为特征提取器。主干网络的权重在 ImageNet [24] 上进行预训练。本文使用一块 24G 显存的 3090 GPU 对模型进行训练，最小批处理大小设置为 4。所有模型都使用 SGD 优化器，动量设置为 0.9，权重衰减为0.000 1。在 Pascal VOC 数据集的基类训练期间，对模型分别进行 240 000、120 000 和 80 000 次迭代的训练，学习率分别为 0.001、0.000 1 和0.000 01。在元微调阶段，模型分别经过 2 000、1 600 和 800 次迭代的训练，学习率分别为 0.001、0.000 1 和 0.000 01。对于 MS COCO 数据集，在训练期间，模型分别进行 56 000、36 000 和 12 000 次迭代的训练，学习率与 VOC 训练设计相同。在元微调过程中，对模型进行 3 000、1 000 和 500 次迭代的微调训练。除此以外，对于引入的参数 λ，根据文献 [12] 的实验结果，在 RPN 和 RCNN 的梯度反传过程中分别设置为 0 和 0.75，元微调训练期间分别设置为 0 和 0.01。

3.2　实验结果及对比

在本节中，将在 PASCAL VOC 和 COCO 数据集上进行实验，并将本文的方法与最先进的方法进行比较。

经过多次随机运行，在表 1中给出了 3 种不同数据分割的 VOC 数据集在新类上的平均评估结果。其中 k-shot 检测是根据 3 种不同的数据拆分设置 k = 1, 2, 3, 5, 10 进行的。本文给出了新类在不同拆分下 IoU 阈值为 0.5（AP50）的 mAP。具体来说，在 k=1 设置下，本文的方法表现远优于 FRCN 和 LSTD这类模型，也优于 Meta RCNN 和 TFA 模型的表现。虽然 MSA-Net 在 split 2 和 split 3 中分别低于 FSDet（2021）和 Repmet，但在 k = 2, 3 设置下，在 3 种不同的分类中，其表现远优于两阶段微调方法，也优于其他加入元学习框架的模型。在 k = 5, 10 设置下的表现也优于其他模型，虽然检测精度提升没有 k=2, 3 设置下明显。可以看出，在同样的评估体系下，本文提出的 MSA-Net 在 PASCAL VOC 数据集的新类上表现均优于目前的一些先进方法，这证明了改进方法的有效性。但由于 PASCAL VOC数据集的数据量仅包括大约 5 000 张图像数据，对于目标检测任务来说，该数据集图像数量过少，仅能作为其中一个评估维度来评估模型，并不具有绝对性。

表 1　在 VOC 2007 测试集上的少样本目标检测表现

Tab. 1　Few-shot object detection performance on VOC 2007 test set

Model	Split 1					Split 2					Split 3					Mean
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10	
FRCN[22]	9.9	15.6	21.6	28.0	35.6	9.4	13.8	17.4	21.9	29.8	8.1	13.9	19.0	23.9	31.0	19.93
Deformable-DETR[25]	5.6	13.3	21.7	34.2	45.0	10.9	13.0	18.4	27.3	39.4	7.3	16.6	20.8	32.2	41.8	23.17
RepMet[25]	26.1	32.9	34.4	38.6	41.3	17.2	22.1	23.4	28.3	35.8	**27.5**	31.1	31.5	34.4	37.2	30.79
FSRW[20]	14.8	15.5	26.7	33.9	47.2	15.7	15.3	22.7	30.1	40.5	21.3	25.6	28.4	42.8	45.9	34.08
Meta RCNN [11]	19.9	25.5	35.0	45.7	51.5	10.4	19.4	29.6	34.8	45.4	14.3	18.2	27.5	41.2	48.1	31.10
TFA[9]	25.3	36.4	42.1	47.9	52.8	18.3	27.5	30.9	34.1	39.5	17.9	27.2	34.3	40.8	45.6	34.71
LSTD[19]	8.2	11.0	12.4	29.1	38.5	11.4	13.8	15.0	15.7	31.0	12.6	18.5	25.0	27.3	36.3	20.39
FSDet[20]	24.2	35.3	42.2	49.1	57.4	**21.6**	24.6	31.9	37.0	45.7	21.2	30.0	37.2	43.8	49.6	36.72
MSA-Net	**26.4**	**41.5**	**47.6**	**49.7**	**58.9**	17.8	**26.6**	**35.4**	**38.1**	**46.5**	21.4	**36.1**	**39.6**	**45.6**	**49.9**	**38.74**

MS COCO 数据集包括 80 个类别，大约 12 万张图像，与 Pascal VOC 数据集相比，检测的图像更加复杂多样，因此实验结果也更具说服力。表 2 展示了 MSA-Net 在 MS COCO数据集的表现。可以看出，在 $k=10$ 设置下，MSA-Net 模型的所有度量标准均优于其他方法。在 $k=30$ 设置下，虽然基于文字和图像的多模态模型 SRR-FSD 在 AP50 的度量标准下检测精度更高，但在更严格的度量标准 AP75 下，检测精度远低于 MSA-Net。这表明 MSA-Net 模型通过改进的模块，可以有效地缓解RPN 错误的区域建议的问题，从而产生更准确的检测结果。在 k=30 设置下，采用 Transformer架构的 Deformable DETR[28]在度量标准 AP 下检测精度略高于 MSA-Net，但其他度量标准以及最终的平均检测精度远低于 MSA-Net。除此以外，MSA-Net 的平均检测精度在所有指标上都高于其他现有的方法如 FSCE 和 MPSR 模型，这意味着 MSA-Net 在少样本情况下具有很强的鲁棒性和泛化能力。

4　消融实验

模块的有效性：对提出的模块在 VOC 2007 test set split 1 数据集上进行全面的消融研究，以验证设计的有效性。所有结果均为 5 次随机运行的平均值。表 3 列出了所有的结果。可以看出，

表 2　在 MS COCO 测试集上的少样本目标检测表现

Tab. 2　Few-shot object detection performance on MS COCO test set

Model	k=10			k=30			Average		
	AP	AP50	AP75	AP	AP50	AP75	AP	AP50	AP75
Faster RCNN[20]	5.5	10.0	5.5	7.4	13.1	7.4	6.45	11.55	6.45
Meta-YOLO[20]	5.6	12.3	4.6	9.1	19.0	7.6	7.35	15.65	6.1
Meta Det[10]	7.1	14.6	6.1	11.3	21.7	8.1	9.2	18.15	7.1
Meta RCNN[11]	8.7	19.1	6.6	12.4	25.3	10.8	10.55	22.2	8.7
TFA[9]	9.1	17.1	8.8	12.1	22.0	12.0	10.6	19.55	10.4
MPSR[16]	9.8	17.9	9.7	14.1	25.4	14.2	11.95	21.65	11.95
SRR-FSD[26]	11.3	23.0	9.8	14.7	**29.3**	13.5	13.0	26.15	13.85
FSCE[27]	11.1	—	9.8	15.3	—	14.2	13.2	—	12.0
Deformable-DETR[28]	11.7	19.6	12.1	**16.3**	27.2	16.7	14.0	23.4	14.4
MSA-Net	**14.5**	**25.1**	**15.2**	16.1	27.6	**16.9**	**15.3**	**26.35**	**16.05**

由于缺少数据，原始的 FRCN 模型严重过拟合。具体地，采用逐步渐进的方式探索每一个模块的效果。(1)在加入梯度反传解耦机制后，模型效果在基类上提升了 12.3%，在新类上有 4.0% 的提升，这表明 RPN 和 RCNN 模块的解耦，提高了模型的泛化能力，这一结果与我们提出的观点一致。(2)在加入注意力蒸馏模块后，模型在新类的效果有显著的提升(7.3%～11.5% mAP)，这表明这一蒸馏模块利用了有限的数据中的有效信息，缓解了过拟合现象。(3)在单独加入多尺度注意力机制后，模型相较于单独加入注意力蒸馏模块的提升较小(5.1%～9.8% mAP)，但同时加入注意力蒸馏模块和多尺度注意力机制模块对模型的提升十分显著，将近 20% mAP，这表明模型可以通过学习不同尺寸的相同特征，增强对特征细节的学习，提升检测精度。(4)从总体上看，加入改进的模块后，模型在新类上的检测精度达到平均 22.6% 的提升，在 $k=2, 3$ 的设置下，提升接近 26%，可以看出，在少样本设置下，改进的模型泛化能力显著增强，且 3 个改进模块之间的交互并不会对模型检测精度产生负面影响。

表 3　改进的模块对模型 mAP(AP50)的影响

Tab. 3　Effect of improved modules on mAP (AP50)

FRCN	Decoupled-layer	ABD	Multi-scale Attention	Base	Novel				
					1	2	3	5	10
√				56.3	9.9	15.6	21.6	28.0	35.6
√	√			68.6	13.5	22.5	24.7	29.9	40.1
√		√		74.5	17.2	27.9	33.1	35.7	45.3
√			√	75.8	15.0	25.8	28.3	32.9	43.9
√	√	√		74.6	21.5	35.2	44.5	45.9	54.7
√	√		√	72.9	19.0	33.7	41.3	43.2	50.1
√		√	√	79.8	23.4	37.1	45.4	48.6	56.3
√	√	√	√	82.0	26.4	41.5	47.6	49.7	58.9

图 7 所示为采用 MS COCO 数据集训练后的模型在 PASCAL VOC 部分样例图像上测试的结果。可以看出 MSA-Net 模型的检测效果明显优于原始 FRCN 模型两阶段微调后的结果，能够检测出部分小目标物体，检测新类的置信度也有一定程度的提高。

FRCN+ft

MSA-Net+ft

图 7　MS COCO 10-shot 设置下训练的模型在 VOC 2007 数据集上的示例

Fig. 7　Example of a model trained under MS COCO 10-shot setting on VOC 2007 dataset

5　结束语

本文对少样本目标检测任务进行了探索，提出了基于多阶段注意力机制的少样本目标检测模型 MSA-Net。模型引进了梯度解耦模块缓解 RPN 和 RCNN 潜在的矛盾，此外，ABD 模块和 Multi-scale 模块充分利用查询集和支持集的

特征信息提高模型的泛化性能。尽管简单，但本文的方法在少样本目标检测的各种基准上达到了先进水平，并通过消融实验验证了模型的有效性。

参考文献：

[1] 李新叶, 龙慎鹏, 朱婧. 基于深度神经网络的少样本学习综述 [J]. 计算机应用研究, 2020, 37(8): 2241 – 2247.

[2] 南晓虎, 丁雷. 深度学习的典型目标检测算法综述 [J]. 计算机应用研究, 2020, 37(S2): 15 – 21.

[3] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91 – 110.

[4] 瑚琦, 卞亚林, 王兵. 基于改进特征金字塔的小目标增强检测算法 [J]. 光学仪器, 2022, 44(5): 14 – 19.

[5] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//Proceedings of the 34th International Conference on Machine Learning. Sydney: JMLR. org, 2017: 1126 − 1135.

[6] SUNG F, YANG Y X, ZHANG L, et al. Learning to compare: relation network for few-shot learning[C]// Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 1199 − 1208.

[7] SNELL J, SWERSKY K, ZEMEL R. Prototypical networks for few-shot learning[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc. , 2017: 4080 − 4090.

[8] CHEN H, WANG Y L, WANG G Y, et al. LSTD: a low-shot transfer detector for object detection [C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI Press, 2018: 2836 − 2843.

[9] WANG X, HUANG T E, DARRELL T, et al. Frustratingly simple few-shot object detection[C]// Proceedings of the 37th International Conference on Machine Learning. Vienna: JMLR. org, 2020: 920−929.

[10] WANG Y X, RAMANAN D, HEBERT M. Meta-learning to detect rare objects[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 9925 − 9934.

[11] YAN X P, CHEN Z L, XU A N, et al. Meta R-CNN: towards general solver for instance-level low-shot learning[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 9577 − 9586.

[12] FU K, ZHANG T F, ZHANG Y, et al. Meta-SSD: towards fast adaptation for few-shot object detection with meta-learning[J]. IEEE Access, 2019, 7: 77597 – 77606.

[13] QIAO L M, ZHAO Y X, LI Z Y, et al. DeFRCN: decoupled faster R-CNN for few-shot object detection[EB/OL]. (2021 –08 –20). https://arxiv.org/abs/2108.09017.

[14] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 2117 − 2125.

[15] KIM Y, KANG B N, KIM D. SAN: learning relationship between convolutional features for multi-scale object detection[C]//Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018: 316 − 331.

[16] WU J X, LIU S T, HUANG D, et al. Multi-scale positive sample refinement for few-shot object detection[C]//Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020: 456 − 472.

[17] EVERINGHAM M, VAN 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.

[18] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//Proceedings of the 13th European Conference on Computer Vision. Zurich: Springer, 2014: 740 − 755.

[19] XIAO Y, LEPETIT V, MARLET R. Few-shot object detection and viewpoint estimation for objects in the wild[C]//Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020: 192 − 210.

[20] KANG B Y, LIU Z, WANG X, et al. Few-shot object detection via feature reweighting[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 8420 − 8429.

[21] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017:

2961 − 2969.

[22] 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.

[23] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011 – 2023.

[24] DENG J, DONG W, SOCHER R, et al. ImageNet: a lange scale hierarchical image database [C]//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009: 248−255.

[25] ZHU X Z, SU W J, LU L W, et al. Deformable DETR: deformable transformers for end-to-end object detection[EB/OL]. (2021 –3 –18). https://arxiv.org/abs/2010.04159.

[26] KARLINSKY L, SHTOK J, HARARY S, et al. RepMet: representative-based metric learning for classification and few-shot object detection[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 5197 − 5206.

[27] ZHU C C, CHEN F Y, AHMED U, et al. Semantic relation reasoning for shot-stable few-shot object detection[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 8782 − 8791.

[28] SUN B, LI B H, CAI S C, et al. FSCE: few-shot object detection via contrastive proposal encoding[C]// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 7352 − 7362.

（编辑：张　磊）