激光与光电子学进展 先进成像

基于关联和识别的少样本目标检测

贾剑利[1,2,3]，韩慧妍[1,2,3*]，况立群[1,2,3]，韩方正[1,2,3]，郑心怡[1,2,3]，张秀权[1,2,3]

[1]中北大学计算机科学与技术学院，山西 太原 030051；

[2]机器视觉与虚拟现实山西省重点实验室，山西 太原 030051；

[3]山西省视觉信息处理及智能机器人工程研究中心，山西 太原 030051

摘要 当前基于深度学习的目标检测算法已较为成熟。然而，基于少量样本检测新类仍具有挑战性，因为少样本条件下的深度学习容易导致特征空间退化。现有工作采用整体微调范式在丰富样本的基类上进行预训练，在此基础上构建新类的特征空间。然而，新类基于多个基类隐式地构造特征空间，其结构较为分散，导致基类与新类之间可分性较差。采用对新类和与其相似的基类进行关联再识别的方法进行少样本目标检测。通过引入动态感兴趣区域头，提升模型对训练样本的利用率，基于二者间的语义相似度，显式地为新类构建特征空间。通过解耦基类和新类的分类分支、添加通道注意力模块及增加边界损失函数，提升二者间的可分性。在标准PASCAL VOC数据集上的实验结果表明，所提方法的nAP50均值较TFA、MPSR及DiGeo分别提升10.2、5.4、7.8。

关键词 少样本目标检测；关联和识别；动态感兴趣区域头；通道注意力；边界损失

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP231658

Few-Shot Object Detection Based on Association and Discrimination

Jia Jianli[1,2,3], Han Huiyan[1,2,3*], Kuang Liqun[1,2,3], Han Fangzheng[1,2,3], Zheng Xinyi[1,2,3], Zhang Xiuquan[1,2,3]

[1]*School of Computer Science and Technology, North University of China, Taiyuan* 030051, *Shanxi, China*;

[2]*Shanxi Key Laboratory of Machine Vision and Virtual Reality, Taiyuan* 030051, *Shanxi, China*;

[3]*Shanxi Province's Vision Information Processing and Intelligent Robot Engineering Research Center, Taiyuan* 030051, *Shanxi, China*

Abstract Deep learning-based object detection algorithms have matured considerably. However, detecting novel classes based on a limited number of samples remains challenging as deep learning can easily lead to feature space degradation under few-shot conditions. Most of the existing methods employ a holistic fine-tuning paradigm to pretrain on base classes with abundant samples and subsequently construct feature spaces for the novel classes. However, the novel class implicitly constructs a feature space based on multiple base classes, and its structure is relatively dispersed, thereby leading to poor separability between the base class and the novel class. This study proposes the method of associating a novel class with a similar base class and then discriminating each class for few-shot object detection. By introducing dynamic region of interest headers, the model improves the utilization of training samples and explicitly constructs a feature space for new classes based on the semantic similarity between the two. Furthermore, by decoupling the classification branches of the base and new classes, integrating channel attention modules, and implementing boundary loss functions, we substantially improve the separability between the classes. Experimental results on the standard PASCAL VOC dataset reveal that our method surpasses the nAP50 mean scores of TFA, MPSR, and DiGeo by 10.2, 5.4, and 7.8, respectively.

Key words few-shot object detection; association and discrimination; dynamic region of interest head; channel attention; margin loss

收稿日期：2023-07-05；**修回日期**：2023-08-09；**录用日期**：2023-08-22；**网络首发日期**：2023-09-19

基金项目：山西省科技重大专项计划“揭榜挂帅”项目(202201150401021)、山西省科技成果转化引导专项(202104021301055)、面向家庭机器人的场景分割及表征方法研究(202303021211153)、中北大学机器视觉与虚拟现实重点实验室研究基金资助项目(447-110103)

通信作者：*hhy980344@163.com

1　引　　言

目标检测是计算机视觉领域的重要任务，旨在定位并识别图像中属于特定类别的目标。近年来，基于深度网络的目标检测算法相比传统算法能够获得更好的检测性能与泛化能力[1]，深度学习方法是一种数据驱动的方法，其研究是建立在大规模数据集的基础上的[2]。当可用的标注数据较少或无法得到大量标注样本时，该类方法难以有效泛化到新类对象。因此少样本目标检测（FSOD）是一个具有重要价值的研究方向，其利用丰富的基类数据提取特征学习相关经验，然后在有限新类注释样本下训练网络。

解决FSOD问题的方法包括元学习[3-5]、度量学习[6]和微调[7-9]等，其中微调方法最为普遍。这些方法多采用双阶段方法，单阶段方法检测精确率较低[10]。Wang等[7]提出了two-stage fine-tuning approach（TFA），通过仅微调Faster R-CNN的边界框分类层和回归层，保持主要组成部分冻结，从而显著提高了检测框架的性能，但TFA易将新类对象错误识别为易混淆的基类。Wu等[8]提出了multi-scale positive sample refinement for few-shot object detection（MPSR），通过数据增强丰富对象尺度，减小样本数据集固有尺度偏差改进TFA，但其正样本细化分支需手动选择。此外，Sun等[9]提出了few-shot object detection via contrastive proposal encoding（FSCE），使用有监督批量对比方法并引入对比建议编码损失缓解错误分类问题。但是，这些方法在微调阶段直接从基类预训练的网络中提取新类的特征，导致新类的特征空间结构不紧凑，类间可分性和检测精度较差。

针对上述问题，本文提出了基于关联和识别的少样本目标检测方法。在PASCAL VOC数据集[11]上进行了大量实验，实验证明所提方法提高了对新类样本的检测精度，优于现有的一些FSOD算法。本文贡献总结如下。

1）提出了基于关联和识别的FSOD两步微调框架，为每个新类构造一个可识别的特征空间。

2）在关联步骤中，基于底层语义相似性构建紧凑的类内结构，实现新类与训练后的相似基类的关联；引入动态感兴趣区域头[12]以更好地利用训练样本，并适应具备更明显特征的高质量样本。

3）在识别步骤中，解耦基类和新类的分类分支，以保证新类与相关联基类之间的可分离性；引入边界损失函数，并在边界框预测环节使用高效通道注意（ECA）[13]模块，以进一步提高类间可分性。

2　相关工作

2.1　Faster R-CNN

Faster R-CNN[14]是一种经典的两阶段目标检测模型，它包含两个主要模块，即用于预测区域建议的深度全卷积网络和Fast R-CNN检测器。Faster R-CNN的基本结构如图1所示。将经过尺寸统一处理的图像输入深度全卷积网络中，提取基础特征图，并将其反馈给区域候选网络（RPN）以获取候选框；所得候选框与基础特征图经感兴趣区域（RoI）池化处理后输入到分类器进行分类，通过边界框回归器得到检测框最终的精确位置。通过预定义的交并比（IoU）阈值进行过滤，为目标分配标签。二分类标签分配范式为

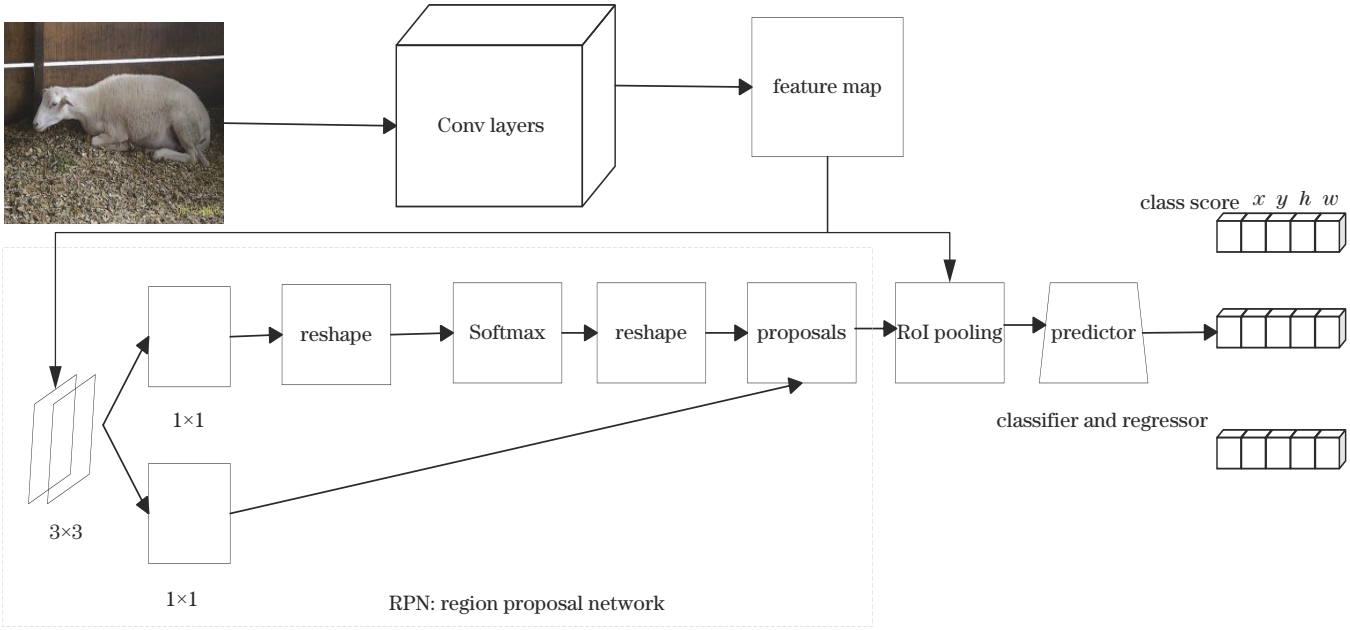

图1　Faster R-CNN的基本结构

Fig. 1　Basic structure of Faster R-CNN

$$c_{\text{label}}=\begin{cases}1, & \max \text{IoU}(\boldsymbol{b},\boldsymbol{G})\geqslant T_{+}\\ 0, & \max \text{IoU}(\boldsymbol{b},\boldsymbol{G})<T_{-},\\ -1, & \text{other}\end{cases}\tag{1}$$

式中：$\boldsymbol{b}$为边界框；$\boldsymbol{G}$为真实值；T_{+}和T_{-}为IoU的正阈值和负阈值（第二阶段中，通常默认为0.5[15]）；1、0、−1分别代表正、负和忽略样本。Faster R-CNN解决了Fast R-CNN需要单独的运算量大的候选区域模块[16]的问题，但网络配置固定，无法充分利用高质量样本，导致RPN无法产生高质量的候选框，从而限制了模型的目标检测性能的提升。

2.2　基于微调的少样本目标检测

基于微调的方法依赖基类信息，采用简单的两阶段训练通道进行少样本目标检测。TFA首先利用丰富样本的基类训练目标检测器，冻结模型其他参数的同时，在由基类和新类组成的平衡训练集上微调检测器的最后几层，其结构如图2所示。将基类图像输入网络后，利用主干网络提取基础特征，并生成相应的区域候选框；候选框与基础特征经过RoI池化处理后进行RoI特征提取，并将提取结果传输至预测层，实现边界框分类和回归。基础训练阶段与微调阶段的主要区别在于输入数据。在微调阶段，新类利用多个与之相似的基类构建自身的特征空间，导致特征空间涵盖多个基类的特征信息，使得空间分布散乱，影响分类性能。

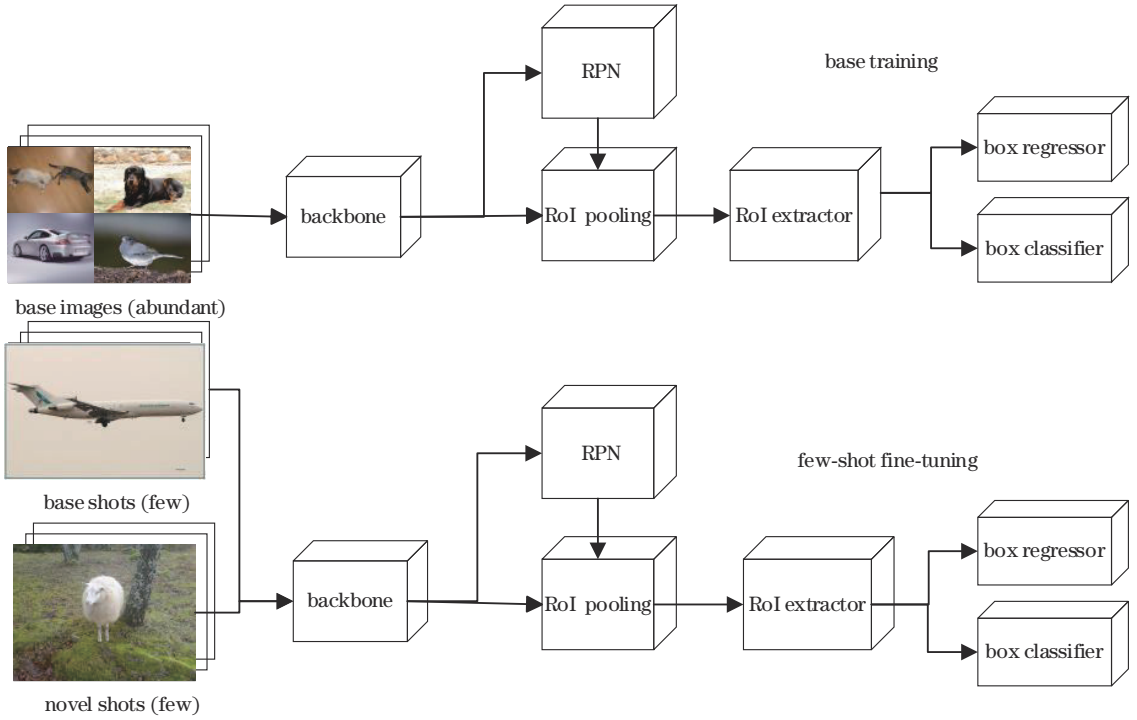

图2　TFA两阶段微调方法的结构

Fig. 2　Structure of TFA two-stage fine-tuning method

3　基于关联和识别的少样本目标检测

目前，少样本目标检测取得了进展，但仍存在利用率低、候选框变化、相似基类混乱等问题。基于TFA框架提出一种新方法，将微调分为关联和识别两步。关联步骤利用动态感兴趣区域头和底层语义的相似性，构建紧凑特征分布。识别步骤解耦基类和新类分支，引入边界损失提升可分性，并添加ECA模块提高分类效率，旨在优化微调阶段性能。

传统的TFA在预训练阶段学习决策边界，将决策空间划分为多个包含不同基类的特征子空间。微调阶段的决策边界如图3所示，新类“摩托车”利用其单一相似基类“自行车”构建特征空间，而新类“牛”可能会利用与之相似的多个基类“羊”和“马”构建自身的特征

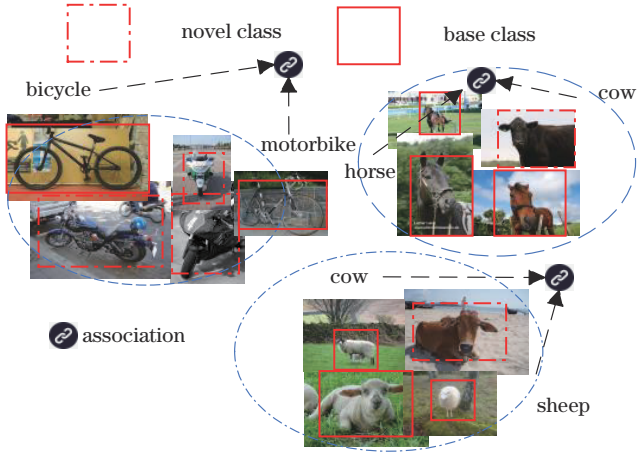

图3　TFA微调阶段的决策边界

Fig. 3　Decision boundary of TFA fine-tuning stage

空间。由于“牛”的特征空间跨越两个基类“羊”和“马”，导致特征空间具有分散的类内结构，“牛”可能被错误识别为“羊”或“马”。为了提升网络模型的类间可分性，将微调阶段分为两个步骤，即关联和识别，如图 4 和图 5 所示。在关联步骤中，基于语义相似性将新类与唯一的基类关联起来。在识别步骤中，引入 ECA 模块，解耦基类和新类的分类分支，增加边界损失函数，最终分离基类和新类。

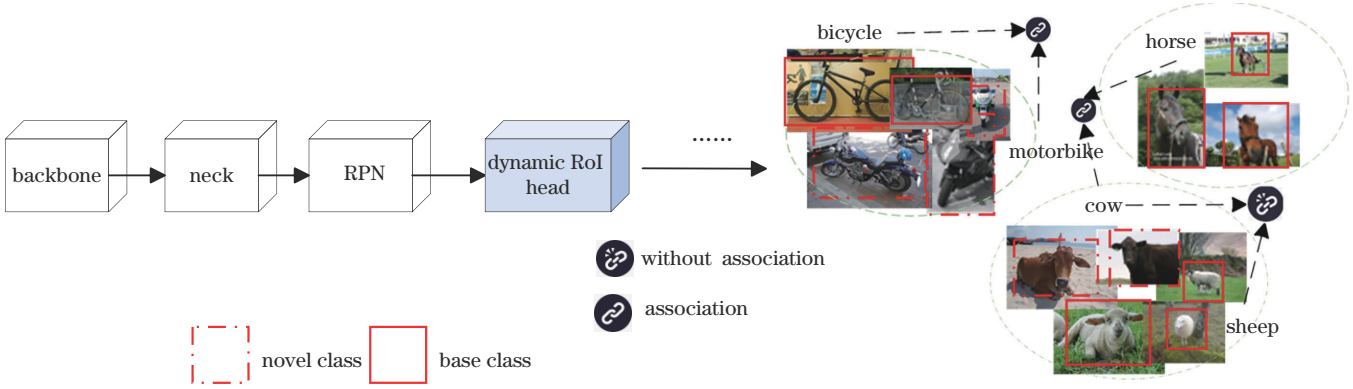

图 4　关联步骤的决策边界

Fig. 4　Decision boundary of association step

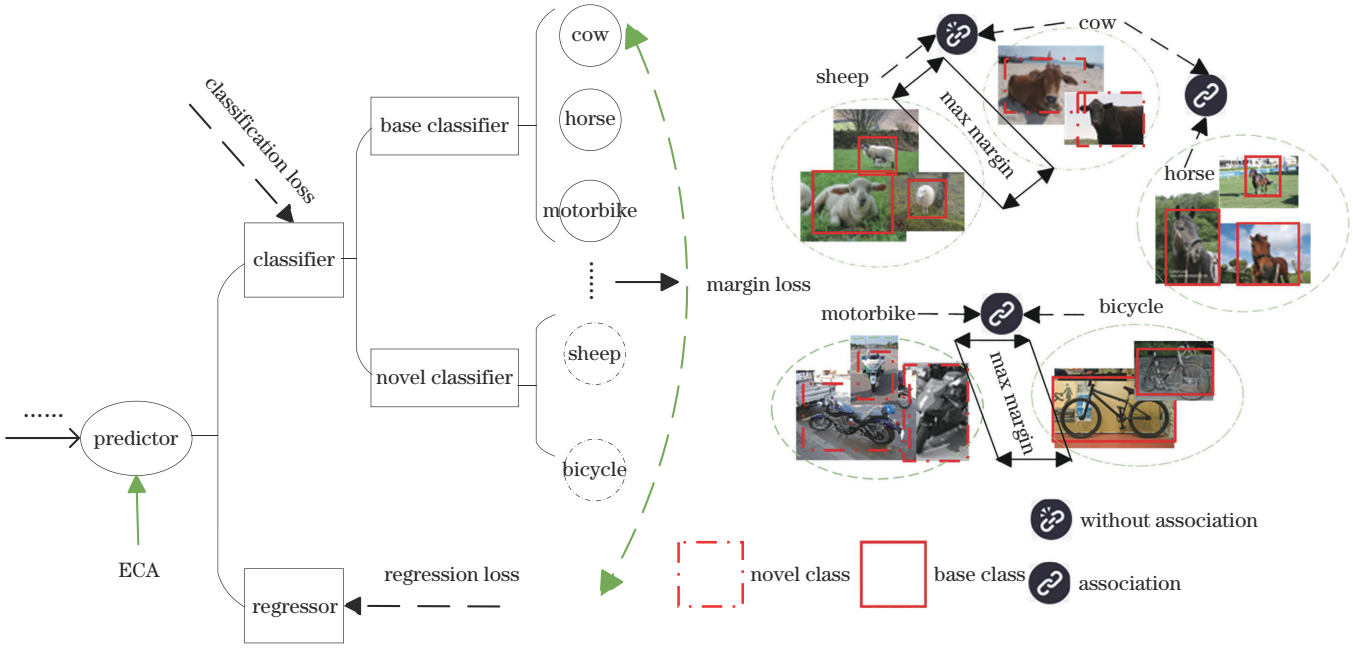

图 5　识别步骤的决策边界

Fig. 5　Decision boundary of discrimination step

3.1　关联新类与基类

在基础训练阶段，将丰富的基类数据集 D^{B} 输入网络模型，通过多轮训练，分类器学习到准确的决策边界。通过将每个新类分配给不同的唯一基类，实现新类间的可分离。为了实现这一目标，根据语义相似性将新类与基类关联，并使用伪标签训练机制将新类和关联基类的特征分布直接对齐，工作步骤如图 6 所示，实现公式为

$$\min_{\boldsymbol{W}_{\mathrm{asso}}^{\mathrm{N}}} \mathcal{L}_{\mathrm{cls}}\left[\boldsymbol{y}_j^{\mathrm{B}}, f\left(\boldsymbol{z}_i^{\mathrm{N}};\ \tilde{\boldsymbol{W}}_{\mathrm{cls}}^{\mathrm{B}}\right)\right]。\tag{2}$$

通过最小化基类伪标签 $\boldsymbol{y}_j^{\mathrm{B}}$ 和基类分类器 $f\left(\cdot;\ \tilde{\boldsymbol{W}}_{\mathrm{cls}}^{\mathrm{B}}\right)$ 预测的分类标签间的分布差异，定义关联阶段分类损失函数。将平衡数据集输入主干网络进行特征提取，通过颈部结构实现不同尺度特征的融合。RPN 提取融合后特征的区域候选框，并将这些候选框传入 RoI 头，输出的 RoI 特征通过全连接层 FC_1 进行处理，利用 $\phi\left(\cdot;\ \tilde{\boldsymbol{W}}_{\mathrm{pre}}^{\mathrm{B}}\right)$ 特征提取器冻结基类预训练所得权重 $\tilde{\boldsymbol{W}}_{\mathrm{pre}}^{\mathrm{B}}$，提取新类样本特征，并将其传输至全连接层 FC_2'。模型的中间结构 $\boldsymbol{z}_i^{\mathrm{N}}=g\left[\phi\left(\boldsymbol{x}_i^{\mathrm{N}};\ \tilde{\boldsymbol{W}}_{\mathrm{pre}}^{\mathrm{B}}\right);\ \boldsymbol{W}_{\mathrm{asso}}^{\mathrm{N}}\right]$ 通过更新关联阶段中的新类权重 $\boldsymbol{W}_{\mathrm{asso}}^{\mathrm{N}}$，实现基类与新类特征分布的对齐，将对齐后的特征传入预测模块，去除回归量，将预测问题简化为简单的分类问题。基类分类器 $f\left(\cdot;\ \tilde{\boldsymbol{W}}_{\mathrm{cls}}^{\mathrm{B}}\right)$ 通过冻结预训练分类阶段基类权重 $\tilde{\boldsymbol{W}}_{\mathrm{cls}}^{\mathrm{B}}$ 进行分类，将生

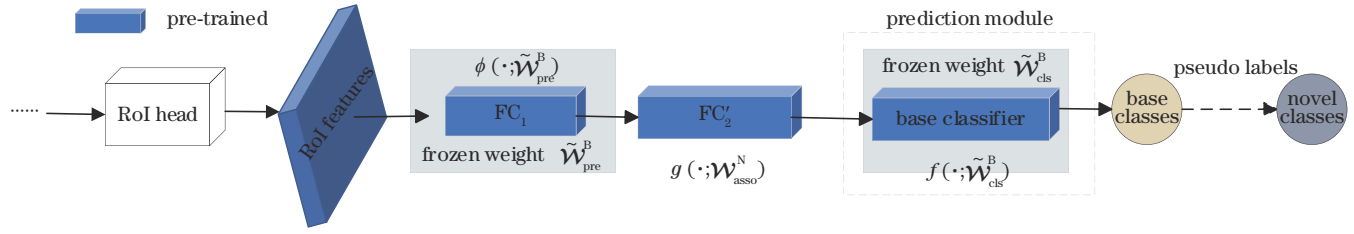

图 6　关联阶段的步骤

Fig. 6　Step of association stage

成的基类伪标签 $\boldsymbol{y}_j^{\mathrm{B}}$ 分配给新类，舍弃与新类关联的基类，最终将新类识别为关联基类。

通过计算基类和新类间的语义相似度将二者关联。给定一个新类 C_i^{N} 和一组基类 C^{B}，使用 WordNet[17] 描述类之间的语义相似度，即计算新类与前 5 个相似基类词汇间的语义相似度，公式为

$$\mathrm{sim}\left(C_i^{\mathrm{N}}, C_j^{\mathrm{B}}\right)=\frac{2\cdot\mathrm{IC}\left[\mathrm{LCS}\left(C_i^{\mathrm{N}}, C_j^{\mathrm{B}}\right)\right]}{\mathrm{IC}\left(C_i^{\mathrm{N}}\right)+\mathrm{IC}\left(C_j^{\mathrm{B}}\right)},\tag{3}$$

式中：函数 LCS(·) 为 WordNet 词法结构中两个类的最小公共子类。以摩托车和牛为例，二者与其最相似基类（自行车和马）间的语义相似度值最大，为 1.0 和 0.795，而牛与羊的相似度值为 0.792。通过在特定语料库 SemCor 中获取信息 i 的词频 $\mathrm{IC}(i)$[18]，对新类与取得最大相似度的基类进行关联，公式为

$$\arg\max_{j\in|C^{\mathrm{B}}|}\mathrm{sim}\left(C_i^{\mathrm{N}}, C_j^{\mathrm{B}}\right)\rightarrow\left(C_j^{\mathrm{B}}\rightarrow C_i^{\mathrm{N}}\right),\tag{4}$$

在确定每个新类的关联基类后，将新类样本 $\boldsymbol{x}_i^{\mathrm{N}}$ 与所分配基类 $C_j^{\mathrm{B}}\rightarrow C_i^{\mathrm{N}}$ 的伪标签 $\boldsymbol{y}_j^{\mathrm{B}}$ 相关联，从而建立新类与基类间的联系。

3.2　识别基类与新类

在关联步骤中，新类与相关联基类的特征分布对齐，使新类具有紧凑的类内分布，并能与其余类分开。然而，这可能会混淆新类和与其关联的基类，影响分类准确率。为此，采取解耦基类和新类的分类分支的策略，并引入边界损失函数，增强类间可分性，其工作步骤如图 7 所示。通过最小化真实标签 $\boldsymbol{y}_i$ 和基类、新类的预测标签 $\left[\boldsymbol{p}^{\mathrm{B}}, \boldsymbol{p}^{\mathrm{N}}\right]$ 间的差异，以达到分类损失函数 $\mathcal{L}_{\mathrm{cls}}$ 的最小化，公式为

$$\min_{\mathcal{W}_{\mathrm{cls}}^{\mathrm{B}}, \mathcal{W}_{\mathrm{cls}}^{\mathrm{N}}}\mathcal{L}_{\mathrm{cls}}\left(\boldsymbol{y}_i, \left[\boldsymbol{p}^{\mathrm{B}}, \boldsymbol{p}^{\mathrm{N}}\right]\right),\tag{5}$$

式中：$\boldsymbol{p}^{\mathrm{B}}$ 和 $\boldsymbol{p}^{\mathrm{N}}$ 为解耦后的基类预测标签和新类预测标签，二者拼接为 $\left[\boldsymbol{p}^{\mathrm{B}}, \boldsymbol{p}^{\mathrm{N}}\right]$。RoI 头输出的 RoI 特征经全连接层 FC_1 传输到冻结基类预训练权重 $\tilde{\mathcal{W}}_{\mathrm{pre}}^{\mathrm{B}}$ 的特征提取器 $\boldsymbol{q}=\phi\left(\boldsymbol{x}_i;\ \tilde{\mathcal{W}}_{\mathrm{pre}}^{\mathrm{B}}\right)$；随后分类提取特征，分为 $g\left(\cdot;\ \tilde{\mathcal{W}}_{\mathrm{origin}}^{\mathrm{B}}\right)$ 和 $g\left(\cdot;\ \tilde{\mathcal{W}}_{\mathrm{asso}}^{\mathrm{N}}\right)$；加载基类预训练后的原始权重 $\tilde{\mathcal{W}}_{\mathrm{origin}}^{\mathrm{B}}$ 并提取基类样本的特征，同时提取解耦后的新类样本的特征，并加载关联阶段的权重 $\tilde{\mathcal{W}}_{\mathrm{asso}}^{\mathrm{N}}$；将这些特征传入基类回归器和基类分类器 $f\left(\cdot;\ \mathcal{W}_{\mathrm{cls}}^{\mathrm{B}}\right)$ 获得基类边界框和分类得分，同时传入新类分类器 $f\left(\cdot;\ \mathcal{W}_{\mathrm{cls}}^{\mathrm{N}}\right)$ 获得新类分类得分。整个过程的表达式为

$$\begin{cases}\boldsymbol{p}^{\mathrm{B}}=f\left[g\left(\boldsymbol{q};\ \tilde{\mathcal{W}}_{\mathrm{origin}}^{\mathrm{B}}\right);\ \mathcal{W}_{\mathrm{cls}}^{\mathrm{B}}\right]\\ \boldsymbol{p}^{\mathrm{N}}=f\left[g\left(\boldsymbol{q};\ \tilde{\mathcal{W}}_{\mathrm{asso}}^{\mathrm{N}}\right);\ \mathcal{W}_{\mathrm{cls}}^{\mathrm{N}}\right]\\ \boldsymbol{q}=\phi\left(\boldsymbol{x}_i;\ \tilde{\mathcal{W}}_{\mathrm{pre}}^{\mathrm{B}}\right)\end{cases}.\tag{6}$$

本文通过增加边界损失函数提升不同类别间的可分性。给定标签为 $\boldsymbol{y}_i$ 的第 i 个训练样本 sample_i，使用余弦相似度计算其属于某类的概率 p_{y_i}，表达式为

$$p_{y_i}=\frac{\tau\cdot\boldsymbol{x}^{\mathrm{T}}\mathcal{W}_{y_i}}{\|\boldsymbol{x}\|\cdot\left\|\mathcal{W}_{y_i}\right\|},\tag{7}$$

式中：$\mathcal{W}_{y_i}$ 为分类器的权重；$\boldsymbol{x}$ 为输入特征；τ 为用于放大梯度的缩放因子。sample_i 的边界损失函数为

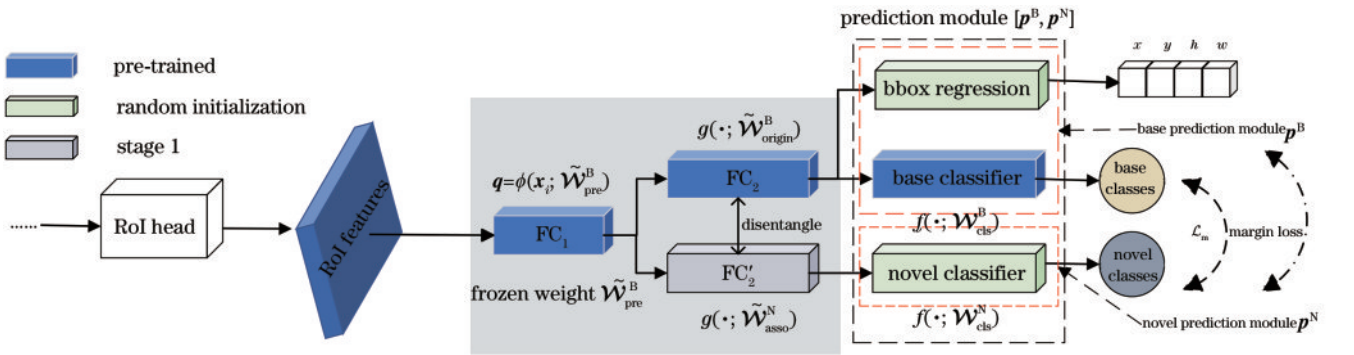

图 7　识别阶段的步骤

Fig. 7　Step of discrimination stage

$$\begin{cases}\mathcal{L}_{\mathrm{m}i}=\sum\limits_{j=1,j\neq y_i}^{C}-\log_{\mathrm{e}}\left[\left(s_{y_i}-s_j\right)^{+}+\varepsilon\right]\\ s_{y_i}=\dfrac{\exp p_{y_i}}{\sum_{j=1}^{C}\exp p_j}\end{cases},\tag{8}$$

式中：s_{y_i}和s_j分别为类C_{y_i}和$C_{j,j\neq y_i}$对应的分类分数；$(\cdot)^{+}$为取正运算；ε是一个很小的数字，避免对数真数取值为0。在少样本学习场景中，由于背景（负）样本在训练样本中的占比较大，需要抑制背景类C_0的边界损失，使新类分类器能够准确识别新类和背景类。通过重加权边界损失函数处理不同类数据不平衡的问题，表达式为

$$\mathcal{L}_{\mathrm{m}}=\sum_{\{i|y_i\in C^{\mathrm{B}}\}}\alpha\cdot\mathcal{L}_{\mathrm{m}i}+\sum_{\{i|y_i\in C^{\mathrm{N}}\}}\beta\cdot\mathcal{L}_{\mathrm{m}i}+\sum_{\{i|y_i\in C_0\}}\gamma\cdot\mathcal{L}_{\mathrm{m}i},\tag{9}$$

式中：α、β、γ分别为平衡基类样本、新类样本和背景样本的边界损失函数的参数。识别阶段采用多任务学习的方式联合优化网络模型，损失函数为

$$\mathcal{L}_{\mathrm{ft}}=\mathcal{L}_{\mathrm{cls}}+\mathcal{L}_{\mathrm{m}}+2\cdot\mathcal{L}_{\mathrm{reg}},\tag{10}$$

式中：$\mathcal{L}_{\mathrm{cls}}$为分类的交叉熵损失；$\mathcal{L}_{\mathrm{reg}}$为边界框回归的平滑L1损失；$\mathcal{L}_{\mathrm{m}}$是边界损失，用于增加不同类之间的差异性，从而提升分类效果。同时回归损失$\mathcal{L}_{\mathrm{reg}}$放大2倍，以平衡分类和回归。

3.3　动态感兴趣区域头

微调过程中，固定的标签分配策略和无法拟合候选框的分布变化的回归损失函数难以训练高质量的检测器，模型性能较差。本文利用动态R-CNN作为RoI头，在训练过程中根据候选框信息自动调整标签分配阈值和回归损失函数的形状，以提高检测器对高质量样本的利用率，其过程如图8所示。

动态标签分配（DLA）过程如图8（a）所示。标签分配策略由式（1）调整为

$$c_{\mathrm{label}}=\begin{cases}1, & \max\mathrm{IoU}(\boldsymbol{b},\boldsymbol{G})\geqslant T_{\mathrm{now}}\\ 0, & \max\mathrm{IoU}(\boldsymbol{b},\boldsymbol{G})<T_{\mathrm{now}}\end{cases},\tag{11}$$

式中：T_{now}为当前阈值。动态平滑L1损失函数（DSL）公式为

$$\mathrm{DSL}\left(\boldsymbol{x},\beta_{\mathrm{now}}\right)=\begin{cases}0.5|\boldsymbol{x}|^2/\beta_{\mathrm{now}}, & |\boldsymbol{x}|<\beta_{\mathrm{now}}\\ |\boldsymbol{x}|-0.5\beta_{\mathrm{now}}, & |\boldsymbol{x}|\geqslant\beta_{\mathrm{now}}\end{cases},\tag{12}$$

式中：$\boldsymbol{x}$为回归标签；β_{now}为控制损失函数使用范围的超参数。随着训练的进行，IoU阈值随着高质量的候选框数目的增多而增大，根据DLA为候选框分配正、负标签，如图8（a）右侧所示。为了适应训练过程中候选框质量变化，对回归损失函数的形状进行了相应的调整，如图8（b）所示，损失函数随回归误差增大而增大，随β的增大而减小，故将β_{now}设为1.0，以提升模型训练的鲁棒性并防止早期网络训练不佳导致的爆炸性损失。

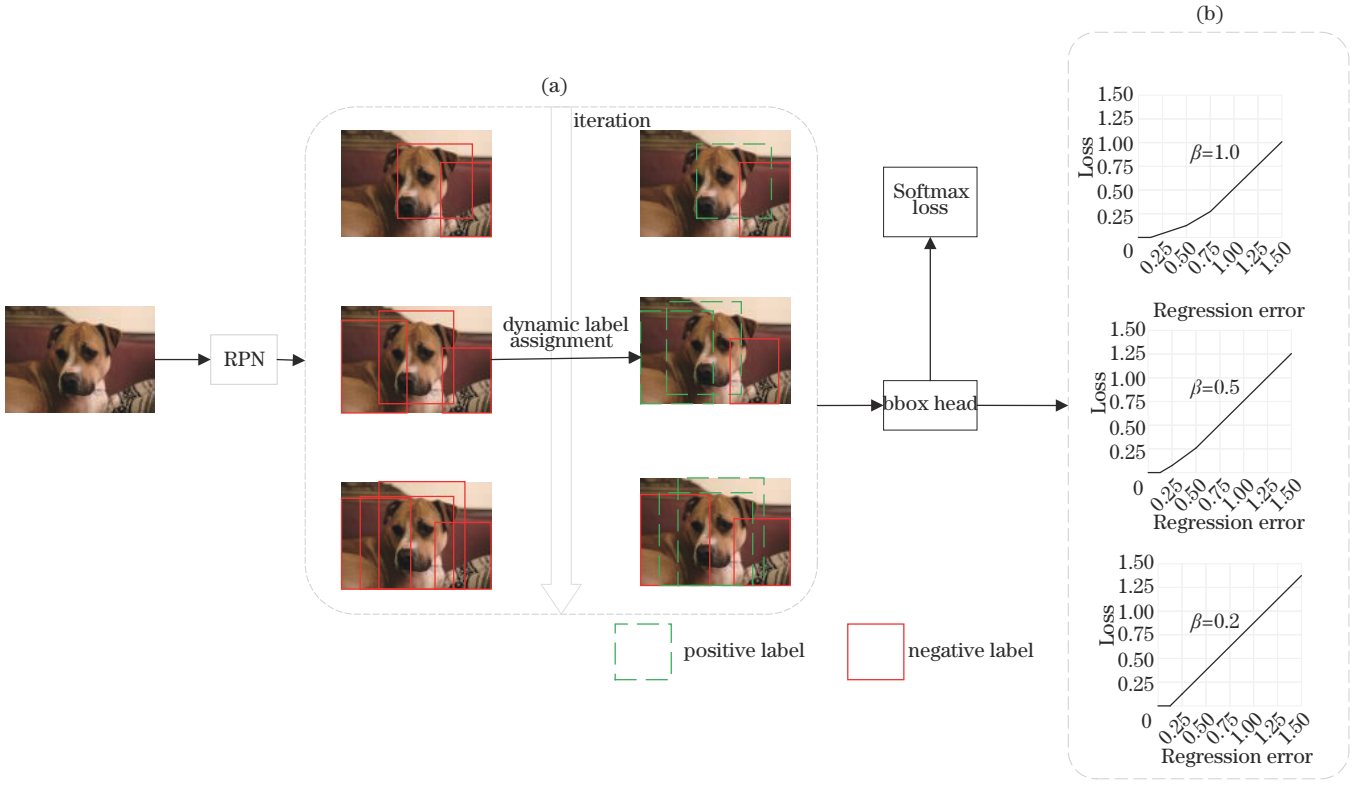

图8　动态R-CNN。（a）DLA；（b）DSL

Fig. 8　Dynamic R-CNN. (a) DLA; (b) DSL

3.4　ECA机制

引入ECA模块以提高网络模型的学习效率[19]。ECA模块利用注意力机制学习输入数据或特征图上不同部分的权重分布，从而减少背景信息对模型的影响，提高模型的识别能力和鲁棒性[20]。如图9所示，输入特征图通过全局平均池化（GAP）聚合特征，进行核

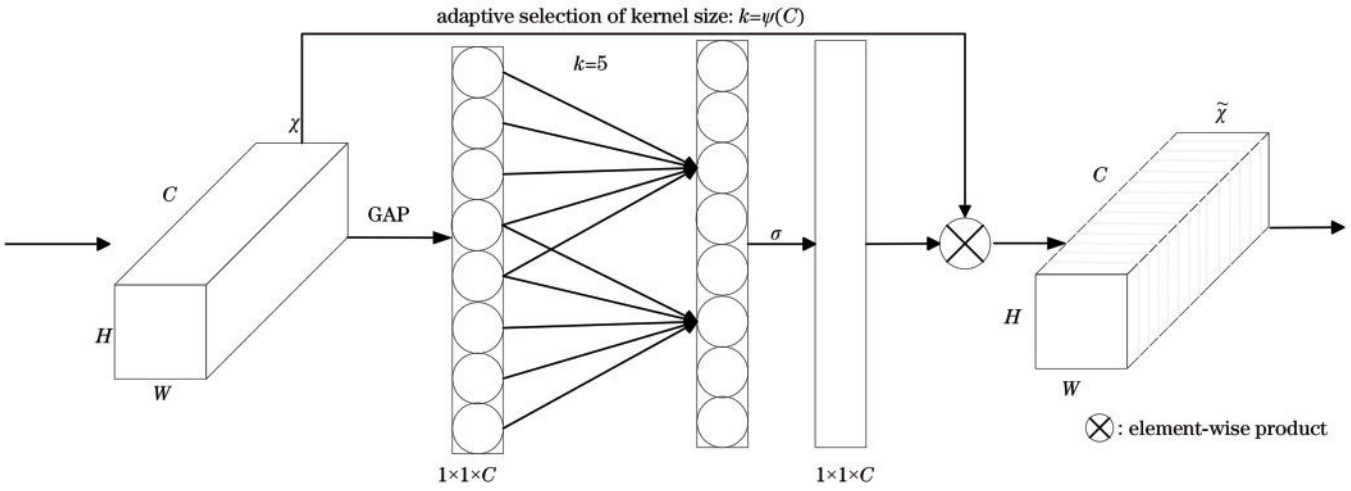

图 9　ECA 模块

Fig. 9　ECA module

大小为 k（由通道维度为 C 的映射函数 $\psi(C)$ 自适应确定）的一维卷积，通过 Sigmoid 激活函数 σ 生成通道权重，将权重和原始特征图对应元素相乘得到最终特征图。这一过程保持了原始数据的维度不变，使得网络模型的性能得到了提升。

4　实验与分析

4.1　数据集

在 PASCAL VOC(07+12)数据集[21]上进行实验，实验严格遵循文献[3,7,9]中使用的数据分割方式和评估标准。PASCAL VOC 包含 20 个对象类别，按照文献[7]的数据分配方式，在 VOC 0712 数据集的训练集和验证集上训练模型，在 VOC 2007 数据集[21]的测试集上进行评估，共 16551 张训练图像和 4952 张测试图像。使用与 TFA 相同的 3 种基类/新类分割方式(Novel Split 1、Novel Split 2、Novel Split 3)，包含 15 个基类和 5 个新类，微调时为每个新类仅提供 K 个注释边界框，其中 K 为 1、2、3、5、10。样本数取决于图像中包含新类对象的数目，所取图像最终包含每个新类的 K 个注释边界框。

4.2　实施细节和评价标准

实验环境为 NVIDIA GeForce RTX 3090 GPU，模型框架为 MMDetection[22]，采用颈部结构为特征金字塔网络[23]及主干网为 ResNet-101[24]的 Faster R-CNN 作为基本模型。在微调阶段，关联和识别步骤使用与 TFA 相同的训练和测试配置。训练过程中，使用水平翻转和随机大小两种数据增强策略，使用随机梯度下降(SGD)优化器优化网络，实验参数如表 1 所示。对训练迭代次数随样本数的变化进行相应的缩放，如表 2 所示。

对于基于 PASCAL VOC 数据集的少样本目标检测，以 nAP50，即 IoU 阈值为 0.5 的新类的平均精度(AP)，作为实验评价指标，表达式为

$$P_{\mathrm{AP}}=\int_0^1 p(r)\mathrm{d}r, \tag{13}$$

表 1　实验参数及其取值

Table 1　Experimental parameters and their values

Parameter	Learning rate	Momentum	Weight decay	Batch size
value	0.001	0.9	0.0001	16

表 2　不同 K 值下的训练迭代次数

Table 2　Number of training iterations under different K values

K	1	2	3	5	10
Number of iterations	4000	8000	12000	16000	20000

式中：p 为 P-R 曲线的纵坐标精确率；r 为横坐标召回率。根据 p 和 r 的值可以绘制 P-R 曲线，P_{AP} 为 P-R 曲线下的面积值。p 和 r 的计算公式分别为

$$p=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}}, \tag{14}$$

$$r=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}}, \tag{15}$$

式中：N_{TP} 为被正确识别的正例样本数，N_{FP} 为被错误识别的负例样本数，N_{FN} 为被错误识别的正例样本数。

4.3　实验结果分析

所提方法是基于 TFA 进行改进的，为了验证所提方法的有效性，进行了与 TFA 的详细比较。表 3 为不同方法在 PASCAL VOC 数据集上的 nAP50，其中最优结果加粗标识。所提方法在任意样本数和分割方式下的性能都明显优于 TFA。在新分割 1、2、3 上，当 K=1、2、3、5、10 时，FSAD 的 nAP50 较 TFA 提高了 1.9 个百分点～18.6 个百分点，且在样本数较少的情况下表现尤其突出。表 3 中的数据表明：所提方法在少样本情况下能有效缓解目标检测问题；与其他少样本方法进行比较时，所提方法在不同样本数下的 nAP50 部分达到最优或次优的效果。图 10 为 FSAD 和 TFA 的预测结果，FSAD 在图像识别精度和边界框回归精度上均优于 TFA，TFA 易将新类对象误识别为易混淆的基

表 3　不同方法在 PASCAL VOC 数据集上的 nAP50

Table 3　nAP50 of different methods on PASCAL VOC dataset　　unit: %

Method	Backbone	Novel Split 1					Novel Split 2					Novel Split 3				
		K=1	*K*=2	*K*=3	*K*=5	*K*=10	*K*=1	*K*=2	*K*=3	*K*=5	*K*=10	*K*=1	*K*=2	*K*=3	*K*=5	*K*=10
LSTD[28]	VGG-16	8.2	1.0	12.4	29.1	38.5	11.4	3.8	5.0	15.7	31.0	12.6	8.5	15.0	27.3	36.3
YOLOv2-ft[29]	YOLO V2	6.6	10.7	12.5	24.8	38.6	12.5	4.2	11.6	16.1	33.9	13.0	15.9	15.0	32.2	38.4
FSRW[3]		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
MetaDet[29]		17.1	19.1	28.9	35.0	48.8	18.2	20.6	25.9	30.6	41.5	20.1	22.3	27.9	41.9	42.9
RepMet[6]	InceptionV3	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
FRCN-ft[29]	FRCN-R101	13.8	19.6	32.8	41.5	45.6	7.9	15.3	26.2	31.6	39.1	9.8	11.3	19.1	35.0	45.1
FRCN+FPN-ft[7]		8.2	20.3	29.0	40.1	45.5	13.4	20.6	28.6	32.4	38.8	19.6	20.8	28.7	42.2	42.1
MetaDet[29]		18.9	20.6	30.2	36.8	49.6	21.8	23.1	27.8	31.7	43.0	20.6	23.9	29.4	43.9	44.1
Meta R-CNN[4]		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
TFA w/fc[7]		36.8	29.1	43.6	55.7	57.0	18.2	29.0	33.4	35.5	39.0	27.7	33.6	42.5	48.7	50.2
TFA w/cos[7]		39.8	36.1	44.7	55.7	56.0	23.5	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8
MPSR[8]		41.7	—	51.4	55.2	61.8	24.4	—	39.2	39.9	47.8	35.6	—	42.3	48.0	49.7
SRR-FSD[30]		47.8	50.5	51.3	55.2	56.8	**32.5**	35.3	39.1	40.8	43.8	40.1	41.5	44.3	46.9	46.4
DiGeo[26]	FRCN-R101	37.9	39.4	48.5	58.6	61.5	26.6	28.9	41.9	42.1	49.1	30.4	40.1	46.9	52.7	54.7
FSCE[9]		44.2	43.8	51.4	61.9	63.4	27.3	29.5	43.5	44.2	50.2	37.2	41.9	47.5	54.6	58.5
Retentive R-CNN[25]		42.4	45.8	45.9	53.7	56.1	21.7	27.8	35.2	37.0	40.3	30.2	37.8	43.0	49.7	50.1
HTRPN[27]		47.0	44.8	53.4	62.9	**65.2**	29.8	32.6	**46.3**	**47.7**	**53.0**	40.1	45.9	**49.6**	**57.0**	**59.7**
FSAD(ours)		**50.5**	**54.7**	**54.6**	57.6	62.2	31.4	**35.5**	39.2	42.5	45.2	**46.1**	**46.3**	47.3	54.8	59.0

图 10　FSAD 与 TFA 预测结果。(a)FSAD;(b)TFA

Fig. 10　Prediction results of FSAD and TFA. (a) FSAD; (b) TFA

类。图 11 为部分算法预测结果图。从图 11 可以看出:MPSR 增加了不同尺度信息,对样本分类准确率产生了一定影响;Retentive R-CNN[25]主要关注基类性能下降问题,未强调新类分类精度提升;DiGeo[26]专注于特征学习,但特征可分性方面存在不足;HTRPN[27]在复杂环境下的检测效果较差;观察表明,FSAD 在识别精度和边界框回归精度上有所提升。表 4 为 FSAD 与部分方法的参数量比较。由表 4 可知,FSAD 的总参数量多于 TFA,少于 DiGeo 和 HTRPN。

4.4　消融实验

对所提方法的各组成部分进行消融研究。首先分析组件的性能贡献,然后展示效果及其工作原理。本文的所有消融结果都是基于 PASCAL VOC 数据集的新分割 1 实现的。

表 5 显示了每个组件的有效性,即关联、解耦和边界损失。加入关联和解耦,实现基类和新类间联系及针对性分离,提升检测精度;边界损失的引入为基类和新类的特征分布划定最大边界,提升类间可分性。当 K=1、3、5 时,相比原始网络,应用这 3 个部分的所提方法的 nAP50 获得了 9.2 个百分点、8.3 个百分点、3.9 个百分点的增益。

表 6 显示关联阶段和识别阶段添加的动态感兴趣

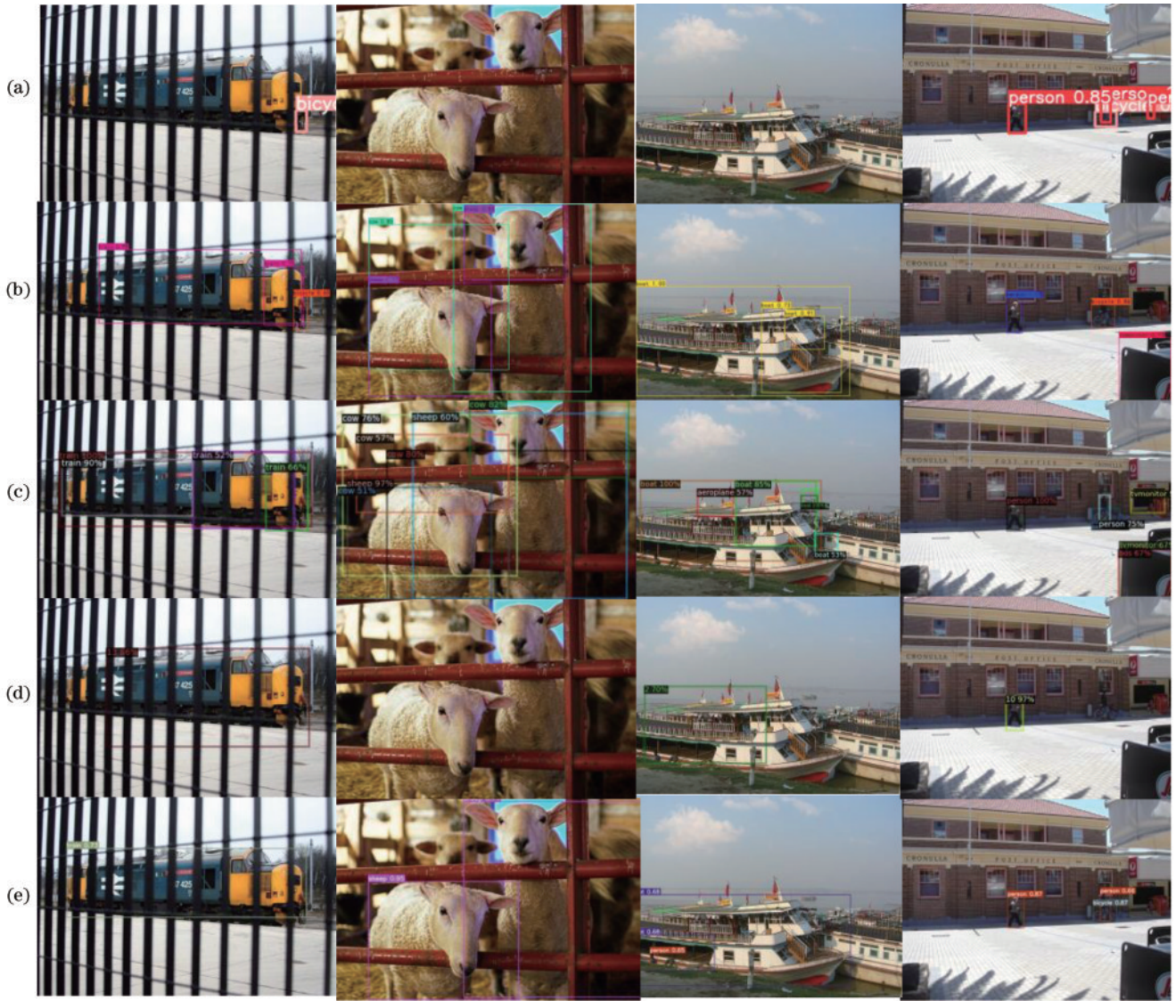

图 11　不同算法的预测结果。(a)MPSR;(b)Retentive R-CNN;(c)DiGeo;(d)HTRPN;(e)FSAD

Fig. 11　Prediction results of differnet algorithms. (a) MPSR; (b) Retentive R-CNN; (c) DiGeo; (d) HTRPN; (e) FSAD

表 4　参数量比较

Table 4　Parameter quantity comparison unit: million

Method	total_params	trainable_params	nontrainable_params
TFA	60.3	0.1	60.2
FSCE	60.3	60.1	0.2
DiGeo	76.4	15.0	61.4
HTRPN	76.5	76.3	0.2
FSAD	60.4	17.9	42.5

表 5　FSAD 不同成分的有效性

Table 5　Effectiveness of different components of FSAD

Association	Disentangling	Margin	nAP50 /%		
			K=1	K=3	K=5
×	×	×	41.3	46.3	53.7
√	×	×	42.4	46.8	55.2
×	√	×	42.4	47.3	54.1
√	√	×	44.9	50.3	56.8
×	×	√	46.3	48.8	56.4
√	√	√	**50.5**	**54.6**	**57.6**

区域头和 ECA 机制的有效性。当 K=1、3、5 时，加入动态感兴趣区域头，模型对样本的利用率提高，提高了对样本的检测精度；加入 ECA 模块后，模型的识别能力和鲁棒性提高。当 K=1、3、5 时，相比原始网络，同时应用这两个模块的网络的 nAP50 获得 7.3 个百分点、5.2 个百分点、3.2 个百分点的增益。

分配策略在关联步骤中具有关键作用。为验证 WordNet 进行语义引导分配的有效性，探索了不同的

表 6　关联和识别阶段模块的有效性

Table 6　Effectiveness of modules in the correlation and recognition stage

Dynamic RoI head	ECA	nAP50 /%		
		K=1	K=3	K=5
×	×	43.2	49.4	54.4
√	×	44.9	51.3	56.0
×	√	46.7	53.0	56.8
√	√	**50.5**	**54.6**	**57.6**

分配策略，结果如表 7 所示。random 表示将基类随机分配给一个新类，human 表示基于人类知识进行的人为分配，visual 表示基于视觉相似性联系基类和新类，top1 和 top2 表示通过式(3)将每个新类分配给与之最相似或第二相似基类的策略。当一个基类被分配给两个不同的新类时(“马”被分配给“鸟”和“牛”)，通过消除重复的方式，将具有最高相似性的基类和新类关联，并重新排序剩余类的相似度，然后选择新的关联。从表 7 可知：相对 random 和 top2，在 top1 下的 nAP50 分别提升了 4.7 个百分点和 3.1 个百分点，表明语义相似度对性能有显著影响；通过消除重复，进一步获得 0.6 个百分点的增益。

表 7　不同分配策略比较(不使用边界损失)

Table 7　Comparison of different allocation strategies (without using margin loss)

Base	bird	bus	cow	motorbike	sofa	nAP50 /%
random	person	boat	horse	aeroplane	sheep	39.6
human	aeroplane	train	sheep	bicycle	chair	44.1
visual	dog	car	horse	person	chair	43.4
top2	dog	car	sheep	tv	diningtable	41.2
top1	horse	train	horse	bicycle	chair	44.3
top1 w/o dup	dog	train	horse	bicycle	chair	**44.9**

对边界损失与 ArcFace[31] 和 CosFace[32] 进行比较，结果如表 8 所示。观察表明，直接应用这两种边界损失会损害性能，且只有应用于新类样本时，才能在一定程度上缓解性能退化。这种应用使得 ArcFace 下的 nAP50 从 37.9% 升至 44.3%，CosFace 下的 nAP50 从 38.9% 升至 44.2%，尽管如此，性能仍比本文的边界损失差 2.0 个百分点。

表 8　不同边界损失性能对比

Table 8　Performance comparison of different margin loss

Margin	nAP50 /%
TFA	41.3
CosFace	38.9
ArcFace	37.9
CosFace(novel)	44.2
ArcFace(novel)	44.3
Ours	**46.3**

通过表 7 可得语义相似度比视觉相似度更有效。将新类实例在基类分类器上的得分预测作为视觉相似度，但是实验发现它有时会产生误导，特别是当一个新类实例与一个基类实例同时出现时，这种共现性欺骗基类分类器，即“马”与“人”、“摩托车”与“人”相似，如图 12 所示，使视觉相似性在数据样本较少的场景下可靠性较低。如表 9 所示，当样本数量增多时，更精确的视觉相似性度量可以减小语义和视觉间的性能差距。

图 12　共存实例，左为语义相似性，右为视觉相似性

Fig. 12　Coexisting instances, left is semantic similarity, right is visual similarity

表 9 视觉和语义相似度的比较

Table 9 Comparison of visual similarity and semantic similarity

Metric	Novel Split 1			Novel Split 2			Novel Split 3		
	K=1	K=3	K=5	K=1	K=3	K=5	K=1	K=3	K=5
Visual	43.3	49.3	56.4	22.5	37.2	39.3	31.8	43.1	50.7
Semantic	**44.9**	**50.3**	**56.8**	**26.1**	**38.5**	**40.1**	**37.1**	**45.0**	**51.5**

5 结 论

针对少样本条件下类别特征空间分散、模型分类能力较差及新类分类精度较低等问题，提出了基于关联和识别的少样本目标检测方法 FSAD。在关联阶段，通过引入动态 RoI 头，充分利用训练样本并构建紧凑的类内结构，利用语义相似性将每个新类和训练后的基类关联，并对齐新类与相关联基类的类内分布。在识别阶段，为了保证类间可分性，解耦基类和新类的分类分支，并增加边界损失函数扩大类间距离。同时，使用 ECA 机制对传入特征向量进行跨通道交互，以提高分类精度。实验结果验证 FSAD 是一种简洁有效的 FSOD 解决方案。

参 考 文 献

[1] 赵菲, 邓英捷. 融合多异构滤波器的轻型弱小目标检测网络[J]. 光学学报, 2023, 43(9): 0915001.
Zhao F, Deng Y J. Light dim small target detection network with multi-heterogeneous filters[J]. Acta Optica Sinica, 2023, 43(9): 0915001.

[2] 李佳男, 王泽, 许廷发. 基于点云数据的三维目标检测技术研究进展[J]. 光学学报, 2023, 43(15): 1515001.
Li J N, Wang Z, Xu T F. Three-dimensional object detection technology based on point cloud data[J]. Acta Optica Sinica, 2023, 43(15): 1515001.

[3] Kang B Y, Liu Z, Wang X, et al. Few-shot object detection via feature reweighting[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Republic of Korea. New York: IEEE Press, 2020: 8419-8428.

[4] Yan X P, Chen Z L, Xu A N, et al. Meta R-CNN: towards general solver for instance-level low-shot learning [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Republic of Korea. New York: IEEE Press, 2020: 9576-9585.

[5] Xiao Y, Lepetit V, Marlet R. Few-shot object detection and viewpoint estimation for objects in the wild[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3): 3090-3106.

[6] Schwartz E, Karlinsky L, Shtok J, et al. RepMet: representative-based metric learning for classification and one-shot object detection[EB/OL]. (2018-06-12) [2023-03-05]. https://arxiv.org/abs/1806.04728.

[7] Wang X, Huang T E, Darrell T, et al. Frustratingly simple few-shot object detection[EB/OL]. (2020-03-16) [2023-03-05]. https://arxiv.org/abs/2003.06957.

[8] Wu J X, Liu S T, Huang D, et al. Multi-scale positive sample refinement for few-shot object detection[M]// Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2020, 12361: 456-472.

[9] Sun B, Li B H, Cai S C, et al. FSCE: few-shot object detection via contrastive proposal encoding[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 7348-7358.

[10] 蔡心悦, 周杨, 胡校飞, 等. 基于超分辨率重建的小目标智能检测算法[J]. 激光与光电子学进展, 2023, 60(12): 1210002.
Cai X Y, Zhou Y, Hu X F, et al. Intelligent detection algorithm for small targets based on super-resolution reconstruction[J]. Laser & Optoelectronics Progress, 2023, 60(12): 1210002.

[11] 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.

[12] Zhang H K, Chang H, Ma B P, et al. Dynamic R-CNN: towards high quality object detection via dynamic training [M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12360: 260-275.

[13] Wang Q L, Wu B G, Zhu P F, et al. ECA-net: efficient channel attention for deep convolutional neural networks [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 11531-11539.

[14] 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.

[15] Girshick R, Radosavovic I, Gkioxari G, et al. Detectron [EB/OL]. [2023-05-03]. https://github.com/facebookresearch/detectron.

[16] 段仲静, 李少波, 胡建军, 等. 深度学习目标检测方法及其主流框架综述[J]. 激光与光电子学进展, 2020, 57(12): 120005.
Duan Z J, Li S B, Hu J J, et al. Review of deep learning based object detection methods and their mainstream frameworks[J]. Laser & Optoelectronics Progress, 2020, 57(12): 120005.

[17] Miller G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39-41.

[18] Lin D. An information-theoretic definition of similarity

[EB/OL]. [2023-05-03]. https://pdfs. semanticscholar.org/cc0c/3033ea7d4e19e1f5ac71934759507e126162.pdf.

[19] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [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: 7132-7141.

[20] 赵斌, 王春平, 付强, 等. 基于深度注意力机制的多尺度红外行人检测[J]. 光学学报, 2020, 40(5): 0504001.
Zhao B, Wang C P, Fu Q, et al. Multi-scale infrared pedestrian detection based on deep attention mechanism [J]. Acta Optica Sinica, 2020, 40(5): 0504001.

[21] Everingham M, Ali Eslami S M, van Gool L, et al. The pascal visual object classes challenge: a retrospective[J]. International Journal of Computer Vision, 2015, 111(1): 98-136.

[22] Chen K, Wang J Q, Pang J M, et al. MMDetection: open MMLab detection toolbox and benchmark[EB/OL]. (2019-06-17)[2023-05-06]. https://arxiv.org/abs/1906.07155.

[23] 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 (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 936-944.

[24] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[25] Fan Z B, Ma Y C, Li Z M, et al. Generalized few-shot object detection without forgetting[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 4525-4534.

[26] Ma J W, Niu Y L, Xu J C, et al. DiGeo: discriminative geometry-aware learning for generalized few-shot object detection[EB/OL]. (2023-03-16) [2023-05-03]. https://arxiv.org/abs/2303.09674.

[27] Shangguan Z Y, Rostami M. Identification of novel classes for improving few-shot object detection[EB/OL]. (2023-03-18)[2023-04-05]. https://arxiv.org/abs/2303.10422.

[28] Chen H, Wang Y L, Wang G Y, et al. LSTD: a low-shot transfer detector for object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 2836-2843.

[29] Wang Y X, Ramanan D, Hebert M. Meta-learning to detect rare objects[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Republic of Korea. New York: IEEE Press, 2020: 9924-9933.

[30] Zhu C C, Chen F Y, Ahmed U, et al. Semantic relation reasoning for shot-stable few-shot object detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 8778-8787.

[31] Deng J K, Guo J, Xue N N, et al. ArcFace: additive angular margin loss for deep face recognition[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: 4685-4694.

[32] Wang H, Wang Y T, Zhou Z, et al. CosFace: large margin cosine loss for deep face recognition[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: 5265-5274.