计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(1): 199−213, 2026

基于跨模块知识蒸馏的3阶段少样本目标检测

王双丞　朱松豪
（南京邮电大学自动化学院、人工智能学院　南京　210023）
（1223055828@njupt.edu.cn）

Three-Stage Few-Shot Object Detection via Cross-Module Knowledge Distillation

Wang Shuangcheng and Zhu Songhao
（*College of Automation & College of Artificial Intelligence*, *Nanjing University of Posts and Telecommunications*, *Nanjing* 210023）

Abstract　Few-shot learning is the mainstream method for few-shot object detection, but it has serious shortcomings. 1) The extreme lack of new class samples leads to biased distribution of new class features; 2) Due to the robustness assumption during the fine-tuning process are not necessarily applicable to new class samples, the feature extraction network is unable to extract unbiased new class sample features. To deal with the above two issues, three-stage fine-tuning few-shot object detection method based on cross-module knowledge distillation is proposed. Firstly, the feature distribution calibration strategy is designed to calibrate the feature distribution of new class during the two-step fine-tuning process. Secondly, the proposed first bias reduction strategy, effectively alleviates the bias estimation problem of weight parameters in the linear probing (the first stage of the fine-tuning process), and the proposed inverse first bias reduction effectively alleviates the over-fitting problem of feature extraction network during overall fine-tuning (the second stage of the fine-tuning process). Finally, the proposed cross-module knowledge distillation strategy is utilized to guide the shallow modules of the model to learn deep features to capture more discriminative new class features. A large number of experimental results show that the proposed three-stage fine-tuning few-shot object detection method effectively improves the accuracy and robustness of few-shot object detection.

Key words　few-shot object detection; three-stage fine-tuning; first bias reduction; inverse first bias reduction; cross-module knowledge distillation

摘　要　少样本学习是少样本目标检测的主流方法，但其存在2个问题：1）新类样本的极度缺乏导致新类特征分布存在偏差；2）由于微调过程中的鲁棒性假设不一定适用于新类样本，因此，特征提取网络无法提取未失真的新类样本特征。为解决上述2个问题，提出基于跨模块知识蒸馏的3阶段少样本目标检测方法。首先，利用特征分布校准策略，校准2步微调过程中新类样本的特征分布；其次，利用提出的首项偏差削减策略，有效缓解线性探测阶段（微调过程的第1阶段）中权重参数的偏差估计问题；然后，利用提出的基于逆首项偏差削减的整体微调策略，有效缓解整体微调过程（微调过程的第2阶段）中特征提取网络的过拟合问题；最后，利用提出的跨模块知识蒸馏策略，引导模型的浅层模块学习深层特征，以便捕获更具辨别力的新类样本特征。大量实验结果表明，所提的3阶段微调的少样本目标检测方法有效提高了少样本目标检测的准确性和鲁棒性。

收稿日期：2025-01-10；**修回日期**：2025-05-26
基金项目：国家自然科学基金项目（52405065）；江苏省研究生科研实践创新项目（SJCX240330）
This work was supported by the National Natural Science Foundation of China (52405065) and Jiangsu Provincial Graduate Research Practice Innovation Project (SJCX240330).
通信作者：朱松豪 (zhush@njupt.edu.cn)

关键词 少样本目标检测;3阶段微调;首项偏差削减;逆首项偏差削减;跨模块知识蒸馏

中图法分类号 TP391.4

DOI: 10.7544/issn1000-1239.202550009 **CSTR:** 32373.14.issn1000-1239.202550009

由于大规模数据集的收集和注释需要大量的人力、物力、财力的投入,因此,少样本学习越来越受到人们关注。为模拟包括罕见疾病、稀有物种、军事目标等数据稀缺情况,研究人员们提出了少样本学习,将从大量有标签的基类样本中学习到的知识,转移至无标签的样本数量有限的新类样本。由于同时包含了目标分类和目标定位的功能,使得现有的少样本目标检测成为可行但具有挑战性的任务,因此,需对少样本目标检测进行更为充分的研究。

为缓解新类和基类间特征分布的差异,少样本目标检测方法通常采用2阶段训练策略。文献[1]验证了基类样本特征可有效表征新类样本特征,最大限度地利用基类和新类样本特征的思想启发了后续研究。首先,补偿基类样本的特征分布,用以防止微调过程中基类样本特征分布的漂移;然后选择与新类样本特征分布具有高度语义相似性的基类样本特征分布进行新类样本特征分布校准,有效促进感兴趣区域网络中语义知识的重用。总之,分布校准(distribution calibration, DC)策略,有效增强了感兴趣区域网络中新类样本目标的定位能力和分类能力。

文献[2]从理论上证明以下结论:对新类检测而言,线性探测(linear probing, LB)性能优于整体微调(overall fine tuning , OFT);对基类检测而言,整体微调性能优于线性探测。因此,线性探测—整体微调结合2者优点,在基类检测和新类检测均有优异性能。详细描述可见文献[2]。尽管线性探测—整体微调策略展现了优异的检测性能,但也存在2个问题:1)在线性探测中,包含足够多新类样本的大规模数据集能够实现检测头权重参数无偏估计,然而,少样本学习中的新类样本极其稀少,使得基类与新类间的特征分布存在明显偏差,从而导致检测头权重参数的错误更新;2)在整体微调过程中,那些与基类存在高度语义相似性的新类往往具有很高的置信度,从而导致特征提取网络的过拟合,而逆首项偏差削减仅对检测头有效,因此,特征提取网络正则化需要借助其他方法加以解决。

针对上述2个问题,本文提出了2步微调策略,如图1所示。1)在2步微调的第1阶段,即线性探测过程中,提出新颖的首项偏差削减(first bias reduction, FBR)策略,用于惩罚均匀分布和预测标签概率分布间KL散度(Kullback-Leibler divergence, KLD)的交叉熵损失,使得预测标签概率分布远离均匀分布,从而滤除低置信度的有偏特征,有效缓解检测头参数更新时错误梯度反向传播的问题。2)在2步微调的第2阶段,即整体微调过程中,提出新颖的逆首项偏差削减(inverse first bias reduction, IFBR)策略,使得预测标签概率分布尽可能接近均匀分布,从而缓解高置信度引起的过拟合问题。实验结果表明,本文所提的2步微调策略,有助于提取未失真的新类样本特征。

由于教师模型在基类上进行预训练,因此在整体微调过程中对新类存在有偏的特征表征。为防止学生模型继承有偏知识,受文献[3]所提方法的启发,本文提出了新颖的跨模块知识蒸馏(cross module knowledge distillation, CMD)模块,以解决学生模型在学习过程中继承教师模型有偏知识的问题。

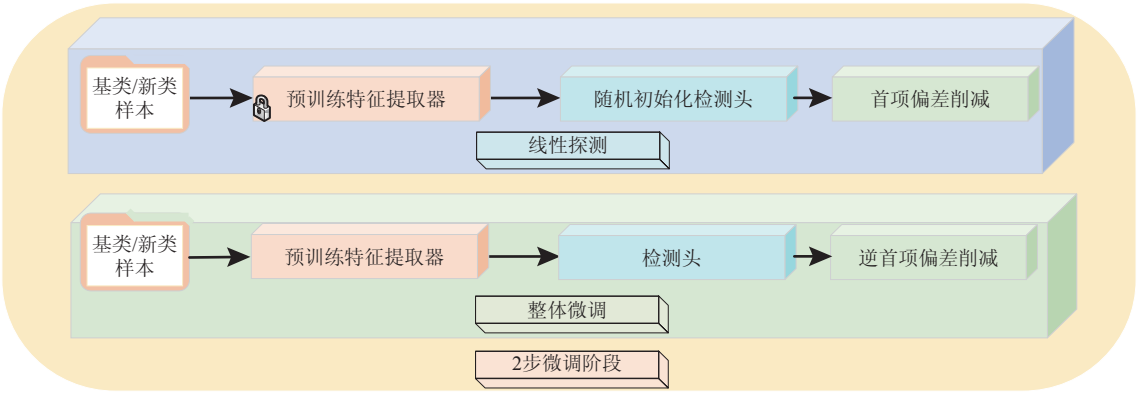


Fig. 1 Framework of the proposed two-step fine-tuning strategy

图1 提出的2步微调策略框架

1 相关工作

1.1 目标检测

目标检测旨在找出所有感兴趣目标，并确定其类别和位置，是计算机视觉领域核心问题之一。现有的目标检测方法通常分为1阶段方法和2阶段方法，且随着深度神经网络的显著进步而不断发展。与同时进行目标定位和目标检测的实时的1阶段方法[4]相比，2阶段方法[5]首先提供建议区域，然后区分和校正每个建议区域，用以实现更高的目标检测精度。

1.2 少样本学习

少样本学习方法旨在利用有限标记的基类图像识别新类图像，主要分为2类：1）基于元学习的少样本学习方法，通过设计特定模型或优化策略实现模型快速自适应[6]；2）基于度量学习的少样本学习方法，设计用于增强鉴别特征的优化嵌入流形[7]。尽管少样本学习和少样本检测具有相似的思路，但相对简单的少样本学习持续推动少样本检测的发展。具体而言，文献[8]通过先验知识转移，生成基于区域建议网络的更为稳健的网络架构，从而有效解决了模型退化问题。文献[9]利用子空间学习对多视图特征进行去噪，用以平衡不同视图间的特征表示，缓解类间差异。文献[10]利用类别分布校准策略，用于将基类的特征分布转移到最相似的新类。少样本学习模型存在当训练样本很少时或类间差异较大，模型知识迁移困难的问题。

1.3 少样本目标检测

少样本目标检测是少样本学习方法在目标检测领域的应用，文献[1]提出了基于微调的少样本目标检测方法，该方法作为开创性的工作，采用2阶段微调，使模型重新适应少量的新类样本，完成知识迁移。文献[11]提出了半监督的少样本目标检测方法，该方法利用教师模型和预训练的少样本目标检测器，通过自适应伪标记，引导学生模型学习。文献[12]提出了判别几何感知模型，旨在改善少样本目标检测任务中的泛化能力。判别几何感知模型通过学习具有类别间分离和类内紧凑性的几何感知特征来提高模型性能。近年来少样本目标检测方法以微调为基本框架，但模型仍然存在过拟合问题。

1.4 分布校准

当训练数据有限时，训练得到的特征分布可能与真实数据分布存在偏差。分布校准通过仅用少数样本来调整有偏的特征分布，从而改善模型的检测性能。文献[13]在探究基类和新类间相似性的基础上，利用基类特征分布的统计信息估计新类的特征分布。文献[14]在探究2阶段学习框架性能瓶颈的基础上，利用提出的分布校准策略，处理不同类别的长尾视觉识别任务。文献[15]通过实验证明了利用分布校准可以改进预训练的回归模型的预测结果。文献[16]利用提出的二进制网络校准潜在特征表示，用以提高特征捕获能力和目标检测精度，该方法将分布校准用于处理回归任务。文献[17]提出建议分布校准策略，通过数据增强手段扩充新类样本数据集，有效解决新类样本的稀缺问题。文献[18]提出保留补偿模块和个性过滤模块，通过特征补偿和过滤缓解新类稀缺造成的类内紧凑性和类间独特性，进而实现知识迁移，而本文的分布校准策略则是通过特征分布的补偿和校准实现知识迁移。文献[19]利用包含双系数的偏移补偿策略调整预训练阶段基类特征分布时，导致了基类特征分布发生扭曲，而本文单系数的漂移补偿策略则能实现基类特征分布的平滑调整。

1.5 微 调

微调作为少样本目标检测基本框架衍生出很多新颖方法。文献[20]提出新颖的三元组网络方法，对具有三元组网络的先验特征空间进行了改进，以学习更具判别性的空间，其中属于同类的特征被聚合，而来自不同类的特征则被分离。文献[21]提出新颖的微调策略，用以学习有监督或自监督的特征表示，实现了与元学习方法相当的检测性能。文献[22]提出基于约束的微调策略，用以缓解灾难性遗忘问题，从而在不增加模型规模的情况下，在新任务上获得优异的检测结果。文献[23]提出基于提示的微调策略，该方法将微策略应用在自然语言处理中，用以解决句子对分类问题。文献[24]提出基于损失削减的微调策略，惩罚低置信度有偏特征相应的损失，从而降低其对模型的影响，用以提高少样本目标检测性能。文献[25]提出新颖的微调策略，调整模型后验概率分布，使其接近均匀分布，有效解决了微调过程中的过拟合问题。目前的微调方法仍然存在缺陷，新类样本稀缺问题和过拟合问题仍然存在。

1.6 知识蒸馏

通过引入与教师网络（复杂但预测精度优越）相关的软目标作为总体损失的一部分，以引导学生网络进行精简、低复杂度、更适合推理部署的训练，实现知识迁移。文献[26]首先选择具有大量语义知识的

开集模型作为教师模型，并对其进行微调以构建学生模型，从而减轻探索新类语义空间的困难；然后，提出后门调整策略，用以缓解由于教师模型提取有偏的特征表示，从而导致学生模型检测性能下降的问题。文献 [27] 提出跨层蒸馏模块，用以引导学生模型的浅层模块学习深层特征，进而缓解新类和基类间显著差异的问题。目前的知识蒸馏无法直接应用在少样本目标检测任务中，由于新类和基类的差异性，教师模型仅对基类进行预训练，学生模型会继承教师模型的有偏新类知识。文献 [28] 提出多层次知识自蒸馏联合多步骤训练的细粒度图像识别，利用判别性强的多粒度特征来提升细粒度图像识别精度。

2 3 阶段微调少目标检测方法

2.1 任务定义

给定一个数据集 D，将其分为基类数据集 D_{b} 和新类数据集 D_{n}，且这 2 个数据集间无重叠 $D_{\mathrm{b}}\cap D_{\mathrm{n}}=\varnothing$。$C_{\mathrm{b}}$ 和 C_{n} 分别为基类和新类的类别数量。每个基类样本都有足够多的有标注的边界框，而每个新类样本只有少数的有标注的边界框。给定一个目标检测数据集 $D=\{[x,y],x\in X,y\in Y\}$，其中 x 表示输入图像，$y=\{(y_m^{\mathrm{cls}},y_m^{\mathrm{reg}})\}$ 表示 x 中的第 m 个目标类别 y_m^{cls} 和对应的边界框 y_m^{reg}。

少样本目标检测模型通常采用 2 阶段训练策略避免过拟合。具体而言，在模型训练的第 1 阶段，即预训练阶段，利用带标签的基类数据集训练模型；在模型训练的第 2 阶段，即微调阶段，仅有检测头在适当的基类样本和新类样本组成的平衡数据集 D_{f} 上进行微调。对于平衡数据集 D_{f}，其类别数为 C，且每个新类中的样本数量 $K\leqslant 10$。

2.2 鲁棒性分析

假设检测头为 $v(\cdot,\boldsymbol{\beta})$（其中 $\boldsymbol{\beta}=(\boldsymbol{\beta}_c)_{c=1}^{C}$ 表示检测头的逻辑回归权重，β_c 表示第 c 个类别的逻辑回归权重），数据集 $D_{\mathrm{f}}=\{(x_m,y_m),1\leqslant m\leqslant M\}$ 共有 C 个类别和 M 个样本，y_m 包括类别标签 y_m^{cls} 和真实框坐标 y_m^{reg}，则正向传播过程可表示为 $\hat{y}_m=v(\boldsymbol{z}_m,\boldsymbol{\beta})=v\{B(x_m,\boldsymbol{\theta})\}$，其中 $\boldsymbol{\theta}$ 表示骨干网络 B 的权重参数。在多元逻辑回归模型中，$\boldsymbol{z}_m$ 到第 c 个类别的分配概率可用式(1)计算得到。

$$Pr(y_m^{\mathrm{cls}}=c|\boldsymbol{z}_m)=\frac{1}{1+\sum_{\bar{c}\neq c,\bar{c}=1}^{C}\mathrm{e}^{\boldsymbol{\beta}_{\bar{c}}^{\mathrm{T}}\boldsymbol{z}_m}}\mathrm{e}^{\boldsymbol{\beta}_c^{\mathrm{T}}\boldsymbol{z}_m}。\tag{1}$$

对应的对数似然函数为

$$\sum_{m\neq 1}^{M}\sum_{c=1}^{C}1[y_m^{\mathrm{cls}}=c]\log\boldsymbol{P}_m^c,\tag{2}$$

其中的 $1[\cdot]$ 以单热编码表示 y_m^{cls}。

在线性探测和整体微调过程中，使用如式(3)所示的交叉熵损失计算分类损失。

$$L_{\mathrm{CE}}=-\frac{1}{M}\sum_{m=1}^{M}[y_m^{\mathrm{cls}}\log\boldsymbol{P}_m],\tag{3}$$

其中的 $\boldsymbol{P}_m=(\boldsymbol{P}_m^c)_{c=1}^{C}$ 表示 $\boldsymbol{z}_m$ 到所有类别的分配概率。此外，使用如式(4)所示的损失函数计算回归损失。

$$L_{\mathrm{reg}}^{\mathrm{roi}}=\frac{1}{M}\sum_{m=1}^{M}R(y_m^{\mathrm{reg}},\hat{y}_m^{\mathrm{reg}}),\tag{4}$$

其中的 $\hat{y}_m^{\mathrm{reg}}$ 表示第 m 个新类样本中预测框的坐标。

假设 S 表示训练样本 X 的子空间，则训练损失可用式(5)计算。

$$L(v(B(X,\boldsymbol{\theta}),\boldsymbol{\beta}),Y)=\|X\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\beta}-Y\|_2^2,\tag{5}$$

其中的 B 表示骨干网络，Y 表示 X 的 1 个热标签。使用式(6)计算骨干网络 B 的训练损失相对于参数 $\boldsymbol{\theta}$ 的梯度。

$$\nabla_{\boldsymbol{\theta}}L(v(B(X,\boldsymbol{\theta}),\boldsymbol{\beta}),Y)=2\boldsymbol{\beta}(Y-X\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\beta})^{\mathrm{T}}X。\tag{6}$$

如果 X' 是与 S 正交子空间的样本，则 X' 的特征在微调过程中不会发生改变。在此情况下，基类特征和新类特征间的不一致将导致骨干网络 B 提取的特征失真：

$$B(\cdot,\boldsymbol{\theta}):\Delta\boldsymbol{z}_m=\nabla_{\boldsymbol{\theta}}L(v(\cdot,\boldsymbol{\beta}),B(\cdot,\boldsymbol{\theta}))X'=2\boldsymbol{\beta}(Y-X\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\beta})^{\mathrm{T}}(X\times X')=0。\tag{7}$$

文献 [2] 从理论上证明了线性探测可以有效解决大规模数据集的微调问题，也即，足够多与基类具有高度语义相似的新类数据能够使得预训练的特征提取网络提取高置信度的特征。然而，在少样本学习中，由于缺乏与基类具有高度语义相似的新类，从而导致预训练的特征提取网络提取低置信度的特征，进而使得检测头的参数更新估计存在偏差。因此，本文提出的首项偏差削减策略旨在滤除低置信度特征，惩罚均匀分布和预测标签概率分布间 KL 散度的交叉熵损失，用以缓解与基类相似度低的新类引起的权重参数偏误估计问题。

2.3 总体框架

本节给出本文所提的基于跨模块知识蒸馏的 3 阶段微调的少样本目标检测的整体框架。如图 2 所示，整体框架包含以下 3 个阶段：1 阶段的基类预训练过程和 2 阶段的微调过程。具体而言：1) 在预训练过程中，利用大量基类样本，获得基类的特征

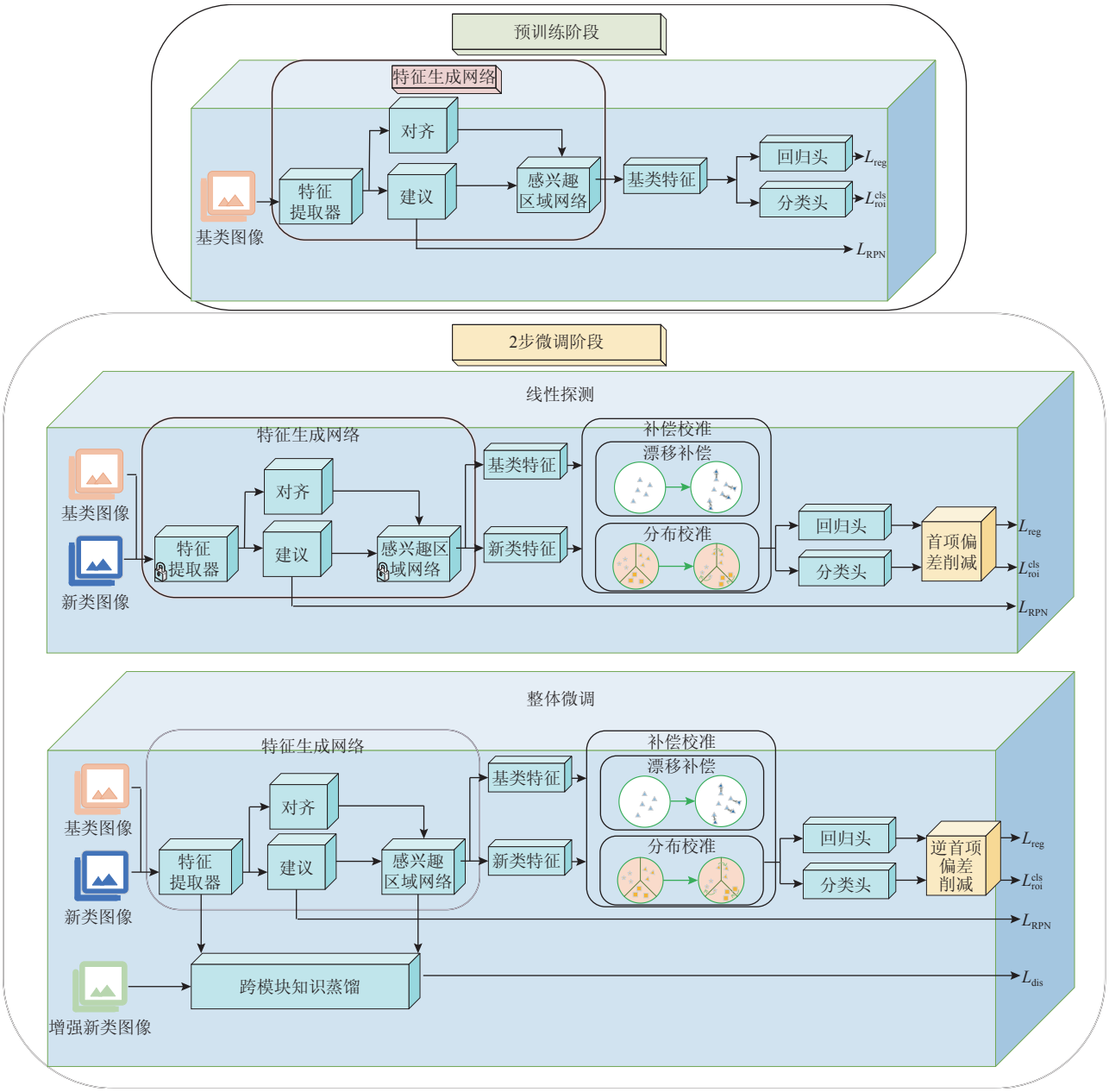


Fig. 2　Framework of three-stage few-shot object detection based on cross-module knowledge distillation

图 2　基于跨模块知识蒸馏的 3 阶段少样本目标检测的框架

分布；2）在微调过程的第 1 阶段，即线性探测过程中，首先使用预训练的特征分布校准新类的特征分布，然后使用提出的首项偏差削减实现权重参数的无偏估计；3）在 2 步微调过程的第 2 阶段，即整体微调过程中，首先使用与线性探测过程相同的分布校准策略，然后使用提出的跨模块知识蒸馏策略引导浅层网络学习深层语义信息，最后使用提出的逆首项偏差削减策略缓解特征提取网络的过参数化问题。

2.4　预训练

1）基类特征标准

如图 2 所示，与其他少样本目标检测方法提出的预训练策略相比，本文提出的预训练策略具有以下特点：所有感兴趣区域特征都被视为正样本（交并比大于等于 0.5），且用于基类特征分布的计算。

2）基类特征分布

特征分布计算过程如下所述。第 1 步，从每个建议区域中提取感兴趣区域特征 $\boldsymbol{x}_i^c \in \mathbb{R}^d$。第 2 步，利用提取的感兴趣区域特征获得所有基类 C_b 的特征分布，且假设每个基类特征分布服从多维高斯分布 $\{\Omega_c = (\boldsymbol{u}_c, \boldsymbol{\sigma}_c^2) | c \in C_b, \boldsymbol{u}_c \in \mathbb{R}^d, \boldsymbol{\sigma}_c^2 \in \mathbb{R}^{d \times d}\}$，其中的 $\boldsymbol{u}_c$ 和 $\boldsymbol{\sigma}_c^2$ 分别是 Ω_c 的均值和协方差，对应式（8）。

$$\begin{cases} \boldsymbol{u}_c = \dfrac{1}{N_c}\sum_{l=1}^{N_c} \boldsymbol{x}_l^c, \\ \boldsymbol{\sigma}_c^2 = \dfrac{1}{N_c-1}\sum_{l=1}^{N_c} (\boldsymbol{x}_l^c - \boldsymbol{u}_c)(\boldsymbol{x}_l^c - \boldsymbol{u}_c)^{\mathrm{T}}, \end{cases} \tag{8}$$

其中的$c \in C_{\mathrm{b}}$表示基类，N_c表示第 c 个基类的正样本数。

2.5 2 步微调策略

与其他 1 步微调策略相比，本文所提的 2 步微调策略具有 2 个特点：

1）由于某些新类和基类间的语义相似性较低，从而导致预训练特征提取网络生成的特征存在偏差。因此，在微调过程的第 1 阶段，即线性探测过程中，利用提出的首项偏差削减策略，调整预测标签概率分布远离均匀分布，从而降低有偏特征的置信度。

2）由于高置信度新类会导致特征提取网络过拟合，因此，微调过程的第 2 阶段，即整体微调，利用提出的逆首项偏差削减策略，调整预测标签概率分布接近均匀分布，从而保证基类检测和新类检测性能的平衡。

2.5.1 线性探测

分布校准的实施过程如下所述。首先，利用高斯分布$\Omega_c = (\boldsymbol{u}_c, \boldsymbol{\sigma}_c^2)$进行基类特征分布的漂移补偿。其次，选择与新类特征分布最相似的C_{b}个基类，进行新类特征的分布校准。

1）漂移补偿

使用少量基类样本和新类样本微调模型参数过程中，采用漂移补偿（drift compensate, DC1）防止添加新类样本导致基类特征分布的漂移，确保新类样本不影响基类的特征分布。

假设所有正基类样本中感兴趣区域的特征分布对应的高斯分布为

$$\{\Omega_c = (\boldsymbol{u}_c, \boldsymbol{\sigma}_c)\}_{c=1}^{C_{\mathrm{b}}}, \tag{9}$$

其中 c 表示基类样本感兴趣区域特征的类别，则采用式（10）进行基类样本特征分布的漂移补偿。

$$\boldsymbol{u}_c = \boldsymbol{u}_c + \alpha \otimes (\boldsymbol{x}_l^c - \boldsymbol{u}_c), \tag{10}$$

其中的 α 为超参数。

2）分布校准

使用补偿分布校准新类的特征分布，也即，利用式（11）所示的最大平均差异（maximum mean discrepancy, MMD）距离度量再生核希尔伯特空间（reproducing kernel Hilbert space, RKHS）中 2 个特征分布的距离，选择与新类感兴趣区域特征$\boldsymbol{x}^{\mathrm{n}}$最相似的 k 个基类，校准新类特征的分布为：

$$\begin{aligned} D_{\mathrm{dist}} &= \{d_c = \|\boldsymbol{x}^{\mathrm{n}} - \boldsymbol{u}_c\|_2, c \in C_{\mathrm{b}}\}, \\ C_{\mathrm{b}}^{\mathrm{cal}} &= \{c | d_c \in \min_k \{D_{\mathrm{dist}}\}, c \in C_{\mathrm{b}}\}, \end{aligned} \tag{11}$$

其中$C_{\mathrm{b}}^{\mathrm{cal}}$表示与新类特征最为相似的基类特征。

利用式（12）校准基类的特征分布。

$$\begin{cases} \boldsymbol{u}_{\mathrm{n}}^* = \boldsymbol{u}_{\mathrm{n}} + \dfrac{1}{k}\sum_{|C_{\mathrm{b}}^{\mathrm{cal}}|} \boldsymbol{u}_c, \\ \boldsymbol{\sigma}_{\mathrm{n}}^* = \boldsymbol{\sigma}_{\mathrm{n}} + \dfrac{1}{k}\sum_{|C_{\mathrm{b}}^{\mathrm{cal}}|} \boldsymbol{\sigma}_c, \end{cases} \tag{12}$$

其中的$\boldsymbol{u}_{\mathrm{n}}$和$\boldsymbol{\sigma}_{\mathrm{n}}$分别表示校准前的新类特征分布均值和方差，$\sum$表示所选 k 个基类的均值或方差。此基础上，利用式（13），获得校准后的新类特征分布。

$$\{\tilde{\boldsymbol{x}}\} = \{\tilde{\boldsymbol{x}} \sim \Omega_{\mathrm{n}} = (\boldsymbol{u}_{\mathrm{n}}^*, \boldsymbol{\sigma}_{\mathrm{n}}^*)\}。 \tag{13}$$

分布校准（distribution calibration, DC）决定了校准特征属性，即基类特征和新类特征间的校准程度越高，基类特征和新类特征在语义空间中的距离就越近，知识共享程度就越高。

3）首项偏差削减

文献 [29] 通过添加对数行列式，消除无偏参数估计中的渐近偏差项$O(n-1)$。文献 [30] 解决了$\det(\boldsymbol{Z}) = 0$时的情况，并提出了一种线性分类器和余弦分类器的近似格式。受该方法启发，本文在线性探测过程中，使用均匀分布和预测标签概率分布间的 KL 散度惩罚交叉熵损失。首项偏差削减策略旨在确保新类预测标签概率分布远离均匀分布，滤除低置信度的偏差特征的同时，确保权重参数的无偏估计，满足训练性能优异检测头的要求。

首项偏差削减的过程可用式（14）表示：

$$L_{\mathrm{roi}}^{\mathrm{cls}} = L_{\mathrm{CE}} - \lambda \frac{1}{M} KL(U_{[0,C]} \| \boldsymbol{P}_m), \tag{14}$$

其中的L_{CE}表示分类检测头的交叉熵损失，λ 表示缩放因子，$U_{[0,C]}$表示 0 和 C 之间的均匀分布。

通过分析式（14），可得 2 个结论：1）如果 λ>0，则预测标签概率分布$\boldsymbol{P}_m$远离均匀分布$U_{[0,C]}$，这与文献 [31] 的结论完全不同；2）如果 λ<0，则预测标签概率分布$\boldsymbol{P}_m$接近均匀分布$U_{[0,C]}$，这与文献 [32] 的结论一致。

基于首项偏差削减和分布校准的线性探测过程有效缓解了权重参数的有偏估计问题。具体线性探测过程如下所述：在 2 步微调的第 1 阶段，即线性探测阶段，利用首项偏差削减和分布校准，对随机初始化的检测头进行再训练和正则化。对于 C-way K-Shot 的少样本学习任务，检测头参数更新迭代过程如式（15）表示。

$$\boldsymbol{\beta}^{*}=\boldsymbol{\beta}+\frac{r_0}{K\times C}\sum_{m=1}^{K\times C}\frac{\partial L(\hat{y}_m,y_m)}{\partial \boldsymbol{r}_m^0}\frac{\partial \boldsymbol{r}_m^0}{\partial \boldsymbol{\beta}}。\tag{15}$$

其中的$\boldsymbol{r}_m^0=\boldsymbol{\beta}B_0(x_m,\boldsymbol{\theta}_0)$，$r_0$表示线性探测阶段检测头的学习率。

2.5.2 整体微调

1)逆首项偏差削减

整体微调阶段，首项偏差削减策略产生不利影响，使得预测标签概率分布偏离均匀分布，这将导致$\hat{y}_m^{\text{cls}}$和y_m^{cls}高度相似，从而导致z_m的高置信度。因此，根据式(14)，提出如图3所示的逆首项偏差削减策略，使得预测标签概率分布$\boldsymbol{P}_m$接近均匀分布。

本文提出的逆首项偏差削减策略类似于标签平滑策略，能够有效减少新类样本对于基类性能的影响，这是因为这2种策略都能使得新类预测标签概率分布接近均匀分布，从而减少新类样本对网络权重参数的影响。同时，这2种策略间的差异也非常明显，本文逆首项偏差削减策略降低了梯度，适用于对参数化模型进行微调。因此，逆首项偏差削减策略优于标签平滑策略。

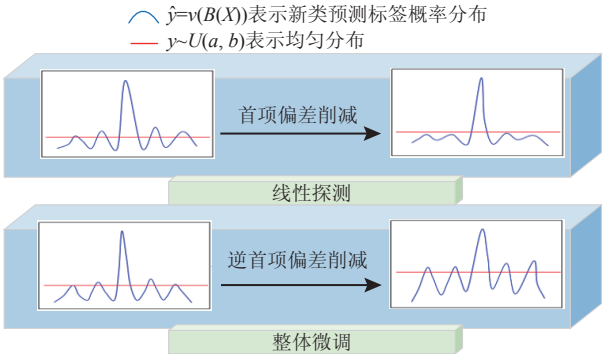


Fig. 3 Examples of the impact of first bias reduction and inverse first bias reduction on predict label probability distribution

图3 首项偏差削减和逆首项偏差削减对预测标签概率分布影响的示例

此外，本文提出的逆首项偏差削减策提供了$L_{\text{roi}}^{\text{cls}}$相对于$\boldsymbol{r}_m=\boldsymbol{\beta}B(x_m,\boldsymbol{\theta})$的梯度，梯度计算过程为

$$\frac{\partial L_{\text{roi}}^{\text{cls}}}{\partial \boldsymbol{r}_m}=\frac{\partial L_{\text{CE}}}{\partial \boldsymbol{r}_m}+\frac{\lambda'}{C+1}(C\boldsymbol{P}_m-\boldsymbol{E}),\tag{16}$$

其中$\boldsymbol{E}$是大小相同且用1填充的向量。从式(16)可知，当$\lambda'<0$且x_m的类别为c时，梯度$\partial L_{\text{roi}}^{\text{cls}}/\partial \boldsymbol{r}_m$减少，这有助于减少新类样本对过度参数化网络权重的影响。

2)跨模块知识蒸馏

在整体微调中，由于基类和新类差异过大导致知识转移困难，因此引入跨模块知识蒸馏方法解决此问题。在教师—学生知识迁移中，教师模型首先在基类数据集上进行预训练；然后，学生模型继承教师模型知识并提取平衡数据集特征表征。然而，由于教师模型仅在基类数据集上进行预训练，从而导致教师模型提取的特征表征存在偏差，进而导致存在偏差的特征表征转移至直接模仿教师模型的学生模型。为缓解这个问题，本文设计了先前学生模型，用于将学生模型的f层特征引入至教师模型的f+1层模块，确保学生模型学习无偏的新类知识。图4给出本文提出的跨模块知识蒸馏框架，具体的跨模块知识蒸馏过程如下。

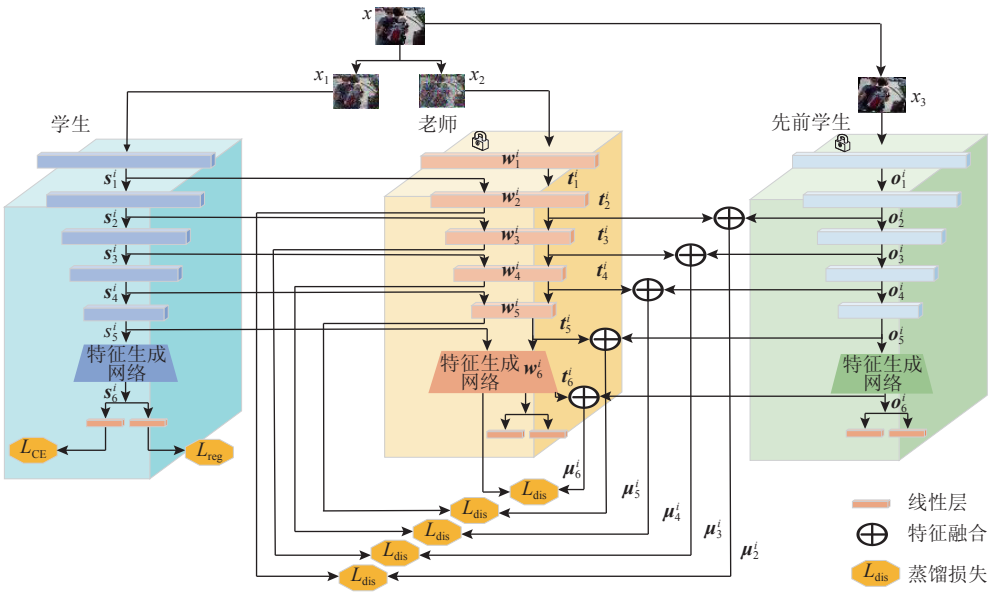


Fig. 4 Architecture diagram of cross-module knowledge distillation

图4 跨模块知识蒸馏的架构图

第 1 步. 将学生模型 N_S、先前学生模型 N_O、教师模型 N_T 中的特征提取器、区域建议网络、感兴趣区域网络统一划分为 F'个模块, 并将由新类样本和基类样本组成的平衡数据集扩展为 3 个部分, 分别作为学生模型、先前学生模型、教师模型的输入。

第 2 步. $\boldsymbol{s}_f^i$, $\boldsymbol{o}_f^i$, $\boldsymbol{t}_f^i$分别表示第 i 次训练迭代后第 f 个模块中的学生模型、先前学生模型、教师模型的特征, 并采用下式对$\boldsymbol{t}_f^i$和$\boldsymbol{o}_f^i$进行特征融合:

$$\boldsymbol{\mu}_f^i = \alpha^i \boldsymbol{o}_f^i + (1-\alpha^i)\boldsymbol{t}_f^i, \tag{17}$$

其中的$\alpha^i = i/T$表示动态权重, T 表示训练迭代次数。

第 3 步. 在第 i 次迭代过程中, 使用式(18)计算跨模块知识蒸馏损失L_{dis}^i:

$$L_{\text{dis}}^i = \begin{cases} \displaystyle\sum_{f=1}^{F'} \| w_{f+1}^i(\boldsymbol{s}_f^i) - \boldsymbol{t}_{f+1}^i \|_2^2 \ (\text{if}(i \leqslant \varpi)), \\ \displaystyle\sum_{f=1}^{F'} \| w_{f+1}^i(\boldsymbol{s}_f^i) - \boldsymbol{\mu}_{f+1}^i \|_2^2 \ (\text{其他}), \end{cases} \tag{18}$$

其中的w_{f+1}^i表示教师模型的第 f+1 个模块, 用于将学生模型$\boldsymbol{u}_{f+1}^i$中的第 f 个模块的特征对齐至同一维度。

第 4 步. 跨模块知识蒸馏的总损失可表示为

$$L = L_{\text{RPN}} + L_{\text{roi}}^{\text{cls}} + L_{\text{reg}} + L_{\text{cal}} + \eta L_{\text{dis}}, \tag{19}$$

其中的L_{RPN}表示由交叉熵损失和回归损失组成的区域建议网络损失, $L_{\text{roi}}^{\text{cls}}$和$L_{\text{reg}}$分别表示感兴趣区域网络的分类损失和回归损耗, L_{cal}和L_{dis}分别表示分布校准损失和跨模块知识蒸馏损失, η 为超参数。在跨模块知识蒸馏过程中, 将学生模型第 f 个模块的特征映射作为教师模型第 f+1 模块的输入进行正向传播, 蒸馏损失的反向传播将完全覆盖学生模型的特征提取网络。

实验结果表明, 图 4 所示的跨模块知识蒸馏策略不仅能引导浅层学习深层语义知识, 且有效缓解学生模型在蒸馏过程中继承教师模型有偏特征表征问题。

基于逆首项偏差削减和跨模块知识蒸馏的整体微调过程有效避免了知识蒸馏中有偏知识转移问题。具体的整体微调过程如下所述: 先使用线性探测进行分布校准, 再利用逆首项偏差削减和跨模块知识蒸馏细化和正则化$B_0(x_m, \boldsymbol{\theta}_0)$和$v^*(\cdot, \boldsymbol{\beta}^*)$:

$$\hat{y}^* = v^*(B_0(x_m, \boldsymbol{\theta}_0), \boldsymbol{\beta}^*)。 \tag{20}$$

整体微调过程的损失为

$$L(\hat{y}^*, y) = \frac{1}{C \times K} \sum_{k=1}^{K \times C} L(\hat{y}_m^*, y_m), \tag{21}$$

其中的y_m表示新类样本 m 的真实标签信息, $\hat{y}_m^*$表示新类样本 m 的预测边界框信息和预测类别信息。

此外, 在整体微调过程中, 特征提取网络参数$B_0(\cdot, \boldsymbol{\theta}_0)$的迭代更新过程为

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + \frac{\alpha_1}{K \times C} \sum_{m=1}^{K \times C} \frac{\partial L(\hat{y}_m^*, y_m)}{\partial \boldsymbol{r}_m^1} \frac{\partial \boldsymbol{r}_m^1}{\partial \boldsymbol{\theta}_0}, \tag{22}$$

其中的α_1表示模型在整体微调过程中的学习率, $\boldsymbol{r}_m^1 = \boldsymbol{\beta}^* B_0(x_m, \boldsymbol{\theta}_0)$。检测头参数的更新过程为

$$\hat{\boldsymbol{\beta}}^* = \boldsymbol{\beta}^* + \frac{r_1}{K \times C} \sum_{m=1}^{K \times C} \frac{\partial L(\hat{y}_m^*, y_m)}{\partial \boldsymbol{r}_m^1} \frac{\partial \boldsymbol{r}_m^1}{\partial \boldsymbol{\beta}^*}, \tag{23}$$

其中的r_1表示整体微调过程中检测头的学习率。

2.6 损失函数

如式(24)所示, 3 阶段微调过程的损失函数包含 3 个部分, 分别是: 1) 预训练阶段的损失, 包括L_{RPN}, $L_{\text{roi}}^{\text{cls}}$, L_{reg}; 2) 线性探测阶段的损失, 包括$L_{\text{roi}}^{\text{cls}}$, L_{reg} , L_{cal}; 3) 整体微调阶段的损失,包括 L_{RPN}, $L_{\text{roi}}^{\text{cls}}$, L_{reg} , L_{cal}, L_{dis}。

$$L = \begin{cases} L_{\text{RPN}} + L_{\text{roi}}^{\text{cls}} + L_{\text{reg}}(\text{预训练}); \\ L_{\text{cal}} + L_{\text{roi}}^{\text{cls}} + L_{\text{reg}}(\text{线性探测}); \\ L_{\text{RPN}} + L_{\text{roi}}^{\text{cls}} + L_{\text{reg}} + L_{\text{cal}} + \eta L_{\text{dis}}(\text{整体微调})。 \end{cases} \tag{24}$$

3 实 验

3.1 数据集

PASCAL VOC 和 MS COCO 是用于评估少样本目标检测方法性能的最广泛使用的公共数据集。与文献 [1] 类似, 首先, 将 PASCAL VOC 数据集分为 3 组, 每组均包含 15 个基类和 5 个新类; 然后, 将每组中 1, 2, 3, 5, 10 个新类样本的平均精度均值(mean average precision, MAP)作为模型的性能评估指标, MAP50, MAP75 分别对应交并比为 0.5 和 0.75 时的平均精度均值。首先, 将 MS COCO 数据集中的所有类别分为 60 个基类和 20 个新类; 然后, 分别计算每个新类的 1, 2, 3 个样本的 MAP, MAP50, MAP75; 最后, 将所有新类类别的 1, 2, 3 个样本的 MAP, MAP50, MAP75 作为模型的性能评估指标。基线模型[1]是典型的少样本目标检测方法, 其基于 Faster RCNN 同时采用 2 步微调的训练策略: 首先在基类数据集上进行预训练, 然后在新类数据集上进行微调, 完成知识迁移。

3.2 实验细节

本文所提的基于跨模块知识蒸馏的 3 阶段少样本目标检测在 PASCAL VOC 数据集上进行了 30 000 次迭代训练, 每批次训练图像数量设为 16。选用动

量为 0.9 和权重衰减为 10^{-4} 的随机梯度下降优化器；选择交并比大于等于 0.5 的基类样本的特征分布作为基类的特征分布；选择 Faster RCNN 作为检测模型，ResNet-50/ResNet-101 作为骨干网络。输入图像尺寸统一设为 640×640。通过渐进调优等方法进一步优化超参数，使模型在验证集上的性能达到最佳。训练中不使用预训练权重。

3.3 实验结果分析

1）PASCAL VOC 数据集上的实验结果分析

表 1 给出不同方法在 PASCAL VOC 数据集的实验结果。与 TFA 相比，本文方法的检测性能有了显著提升，其中在新类样本为 1 个时，MAP50 分别增加了 21.5 个百分点、24.2 个百分点、25.9 个百分点。与 TFA 的结合联合区域建议网络的基线方法相比，本文方法的检测性能也有显著提升，其中在新类样本为 10 个时，MAP50 分别增加了 11.7 个百分点、16.7 个百分点、12.7 个百分点。与 DK 相比，在新类样本为 5 个时，本文方法的 MAP50 分别超过了 9.1 个百分点、4.9 个百分点、7.6 个百分点。与其他少样本目标检测方法相比，本文方法的检测性能也有一定程度的提高。表 1 所示的实验结果与本文的 4 个构想一致：1）提出的分布校准策略能够有效解决新类样本的数据稀缺问题；2）在少样本学习中，与基类相似度低的新类将会产生有偏特征，而提出的首项偏差削减策略可有效缓解由此引起的权重参数错误更新；3）与基类相似度高的新类将会导致特征提取网络的过拟合，而提出的逆首项偏差削减策略可有效减少新类样本影响，避免基类的检测性能下降；4）提出的跨模块知识蒸馏策略可有效减轻基类和新类间特征分布差异造成的知识转移困难，同时避免了教师模型传递错误知识至学生模型。

Table 1 Detection Results on PASCAL VOC Dataset in Terms of MAP50

表 1 在 PASCAL VOC 数据集的以 MAP50 为评价指标的检测结果 %

方法	第 1 组					第 2 组					第 3 组				
	1 个	2 个	3 个	5 个	10 个	1 个	2 个	3 个	5 个	10 个	1 个	2 个	3 个	5 个	10 个
FSCE [33]	44.2	43.8	51.4	61.9	63.4	23.7	30.6	38.4	38.4	48.5	37.2	41.9	47.5	54.6	58.5
MPSR [34]	41.7	43.1	51.4	55.2	61.8	22.4	29.3	39.2	39.9	47.8	35.6	40.6	42.3	48.0	49.7
VFA[35]	57.7	64.6	64.7	67.2	67.4	41.4	46.2	51.1	51.8	51.6	48.9	54.8	56.6	59.0	58.9
Dc-Net[36]	33.9	37.4	43.7	51.1	59.6	23.2	24.8	30.6	36.7	46.6	32.3	34.9	39.7	42.6	50.7
TFA-CoRPNs [1]	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
TFA [1]	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
KFSOD [37]	44.6		54.4	60.9	65.8	37.8		43.1	48.1	50.4	34.8		44.1	52.7	53.9
Meta Faster-RCNN [38]	43.0	54.5	60.6	66.1	65.4	27.7	35.5	46.1	47.8	51.4	40.6	46.4	53.4	59.9	58.6
FPD [39]	37.7	51.2	59.0	64.7	67.8	28.7	40.0	44.3	50.2	55.5	29.2	48.3	52.1	58.4	62.1
DR[26]	41.0	51.7	55.7	61.8	65.4	30.7	39.0	42.5	46.6	51.7	37.9	47.1	51.7	56.8	59.5
ResNet-101（本文）	**58.3**	**64.7**	**65.3**	**70.9**	**68.7**	**42.4**	**46.7**	**51.3**	**51.5**	**55.7**	**53.6**	**50.4**	**57.9**	**64.4**	**62.9**

注：黑体数值表示最优值，每组取 1，2，3，5，10 个样本。

2）MS COCO 数据集上的实验结果分析

表 2 给出不同方法在 MS COCO 数据集检测结果。与 TFA CoRPNs 结合联合区域建议网络的基线相比，在新类样本数量为 1, 2, 3 时，本文方法的 MAP 分别超过了 2.8 个百分点、3.8 个百分点、3.0 个百分点。与 Meta Faster RCNN 相比，在不同新类样本数量时，本文方法的 MAP 分别超过了 1.1 个百分点、0.8 个百分点、1.3 个百分点。与最新的少样本目标检测方法相比，本文方法的检测性能也有显著提高，表明本文方法具有很强的竞争力。

3）计算复杂度分析

为全面评估本文方法的综合性能，对包括模型参数量、迭代时间（微调阶段）、推理时间在内的模型效率进行评估。由表 3 所示实验结果可得 3 个结论：1）由于引入了跨模块知识蒸馏模块，本文方法的模型参数量相比基线方法有所增加，但相比其他方法，本文方法的模型参数量具有一定优势。2）本文方法的迭代时间具有一定优势。相比 Meta Faster-RCNN 方法，本文方法的迭代时间降低 0.13 s，相比 MPSR 方法，本文方法的迭代时间降低 0.298 s。3）本文方法的推理时间具有一定优势。相比 Meta Faster-RCNN 方法，本文方法的推理时间降低 0.014 s，相比

Table 2 Detection Results on MS COCO Dataset

表 2 在 MS COCO 数据集的检测结果 %

方法	1 个新类样本			2 个新类样本			3 个新类样本		
	MAP	MAP50	MAP75	MAP	MAP50	MAP75	MAP	MAP50	MAP75
MPSR [34]	2.3	4.1	2.3	3.5	6.3	3.4	5.2	9.5	5.1
FsDetView [40]	3.2	8.9	1.4	4.9	13.3	2.3	6.7	18.6	2.9
Hallucination [41]	4.4	7.5	4.9	5.6	9.9	5.9	7.2	13.3	7.4
MPF-Net [42]	4.6	11.4	4.9	5.7	10.9	6.2	7.4	13.5	7.8
TFA CoRPNs [1]	3.4	5.8	3.8	4.6	8.3	4.8	6.6	12.1	6.5
Meta Faster-RCNN [38]	5.1	10.7	4.3	7.6	16.3	6.2	8.3	15.2	8.0
ResNet-101 (本文)	**6.2**	**12.1**	**9.3**	**8.4**	**16.6**	**8.2**	**9.6**	**19.3**	**10.7**

注：黑体数值表示最优值。

Table 3 Computational Complexity on PASCAL VOC Dataset

表 3 在 PASCAL VOC 数据集的计算复杂性

方法	模型参数量	迭代时间/s	推理时间/s
FSCE [33]	308×10^6	1.061	0.079
MPF-Net [42]	335×10^6	0.908	0.083
TFA [1]	231×10^6	0.979	0.081
Meta Faster-RCNN [38]	441×10^6	1.012	0.092
MPSR [34]	636×10^6	1.180	0.110
ResNet-101 (本文)	341×10^6	0.882	0.078

MPSR 方法，本文方法的推理时间降低 0.032 s。综上所述，本文方法虽然在模型参数量上有所增加，但模型的迭代时间和推理时间优于其他方法。

3.4 消融研究

1)分布校准的消融实验

图 5 给出基线方法的新类样本与本文方法的加入分布校准的新类样本的边界框定位能力的实验对比，其中横轴表示建议框与真实框的交并比，纵轴表示建议框的比例。此外，方块虚线表示预训练后，基类建议框的比例随交并比变化的情况；菱形块虚线表示微调后，基线方法中新类建议框的比例随交并比变化的情况；原球实线表示 2 步微调后，本文方法中新类建议框的比例随交并比变化的情况。图 5 可知，利用分布校准策略获得的新类的交并比更倾向于基类的交并比，这表明分布校准策略可有效将建议边界框定位能力从基类转移到新类。

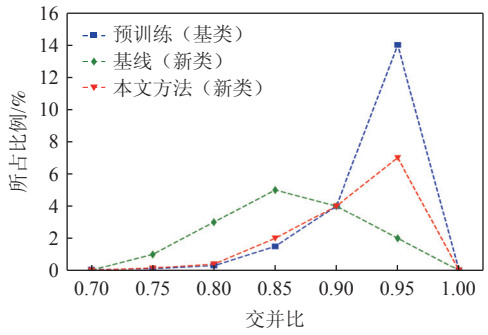


Fig. 5 Proportion of proposal boxes varies with the intersection ratios under different fine-tuning strategies

图 5 不同微调策略下建议框所占比例随交并比变化

图 6 给出了不同微调策略下，建议框与真实框的目标分类置信度随交并比的变化结果。此外，圆球虚线表示预训练后，分类置信度随交并比变化的情况；方块虚线表示基线方法微调后得到的分类置信度随交并比变化的情况；菱形块虚线表示加入本文提出的漂移补偿和分布校准的微调策略后，得到的分类置信度随交并比变化的情况。图 6 可知，漂移补偿和分布校准有效提高了目标的分类置信度，这与本文的以下 2 个构想相一致：1) 在微调过程中，漂移补偿有效缓解了引入新类样本对原有基类特征分布的影响；2) 分布校准有利于将语义知识从基类转移至新类。

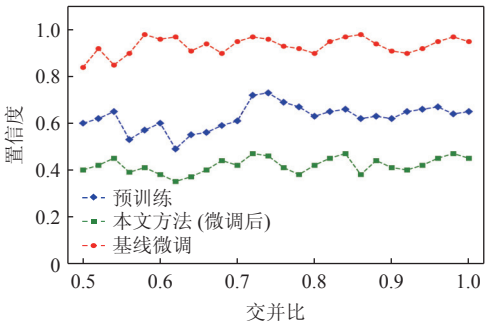


Fig. 6 Confidence of target prediction varies with the intersection ratio curve under different fine-tuning strategies

图 6 不同微调策略下目标预测的置信度随交并比变化

由表 4 所示的 PASCAL VOC 数据集的消融实验结果可知，本文提出的分布校准策略对模型性能有一定程度的提升。在新类样本数量为 2 的条件下，在基线中添加漂移补偿(α=0.5)后，3 组的 MAP50 分

Table 4 Ablation Experiments on PASCAL VOC Dataset

表 4 PASCAL VOC 数据集上的消融实验 %

索引	TFA	DC1	DC2	第 1 组					第 2 组					第 3 组				
				1 个	2 个	3 个	5 个	10 个	1 个	2 个	3 个	5 个	10 个	1 个	2 个	3 个	5 个	10 个
1	√			36.8	39.1	43.6	48.7	50.0	20.2	25.0	29.4	32.5	34.0	27.7	31.6	38.5	44.7	48.2
2	√	α=0.1		38.4	42.6	45.1	49.5	51.2	21.7	26.0	30.9	34.3	35.5	28.6	32.2	38.9	45.4	49.3
3	√	α=0.3		40.5	45.8	47.3	51.1	53.8	23.2	27.4	31.5	36.1	37.9	29.4	33.5	39.2	47.1	50.6
4	√	α=0.5		**42.7**	**49.2**	**50.3**	**53.2**	**54.9**	**25.0**	**29.9**	**33.3**	**37.3**	**39.5**	**31.1**	**34.7**	**39.9**	**48.6**	**51.1**
5	√	α=0.7		41.1	46.9	48.5	52.6	53.4	23.7	27.6	32.2	36.4	38.2	29.8	33.9	39.7	47.5	50.8
6	√	α=1.0		39.0	43.1	45.7	50.2	52.3	22.4	26.9	31.5	34.8	36.3	29.1	32.7	39.5	45.9	50.3
7	√	α=0.5	√	**44.6**	**51.7**	**52.7**	**56.6**	**58.1**	**27.6**	**33.7**	**38.1**	**40.2**	**42.9**	**35.6**	**39.3**	**44.3**	**51.7**	**53.4**

注：黑体数值表示最优值，DC1 表示漂移补偿，DC2 表示分布校准，α 表示分布校准的超参数，“√”表示使用该策略。

别达到 49.2%, 29.9%, 34.7%; 在此基础上，添加分布校准后，3 组的 MAP50 分别达到 51.7%, 33.7%, 39.3%; 从表 4 最后 1 行的实验结果可知，在添加了包括漂移补偿和分布校准的分布校准策略后，在新类样本数量为 3 的条件下，3 组的 MAP50 分别达到 52.7%, 38.1%, 44.3%。

2）2 步微调的消融实验

从表 5 中索引 1 和索引 5 的对比实验结果可知，在添加首项偏差削减线性探测后，模型性能在一定程度上得到提升，如当新类样本数量为 1 时，MAP50 分别增加了 1.6 个百分点、3.0 个百分点、2.7 个百分点，这表明首项偏差削减有利于训练性能良好的检测头。从表 5 的对比实验结果可知，在添加了逆首项偏差削减的整体微调后，模型性能再次得到提升，如当新类样本数量为 1 时，MAP50 分别再次增加了 0.6 个百分点、2.0 个百分点、1.1 个百分点，这表明逆首项偏差削减可有效避免特征提取网络的过参数化，并确保基类检测性能不受影响。

Table 5 Ablation Experiments on PASCAL VOC Dataset in Terms of MAP50

表 5 PASCAL VOC 数据集的以 MAP50 为评估指标的消融实验 %

索引	DC	LP-FBR	OFT-IFBR	第 1 组					第 2 组					第 3 组				
				1 个	2 个	3 个	5 个	10 个	1 个	2 个	3 个	5 个	10 个	1 个	2 个	3 个	5 个	10 个
1	√			44.6	51.7	52.7	56.6	58.1	27.6	33.7	38.1	40.2	42.9	35.6	43.3	46.3	53.7	53.4
2	√	λ=0.1		44.9	52.0	53.6	57.5	59.3	28.4	34.1	39.5	41.6	43.5	36.3	43.5	46.7	54.2	54.5
3	√	λ=0.3		45.5	52.4	54.5	58.2	60.6	29.4	34.9	40.8	42.4	45.0	37.2	43.9	47.0	54.5	56.3
4	√	λ=0.5		45.9	52.8	55.6	59.3	61.5	30.1	35.4	41.4	43.6	45.8	37.8	44.5	47.2	54.8	57.1
5	√	λ=0.7		**46.2**	**53.2**	**56.8**	**60.1**	**62.2**	**30.6**	**36.1**	**42.7**	**44.0**	**46.2**	**38.3**	**44.8**	**47.5**	**55.1**	**58.6**
6	√	λ=1.0		45.6	52.7	55.1	58.9	61.1	29.9	35.2	41.1	42.9	45.3	37.5	44.0	47.1	54.6	56.9
7	√	λ=0.7	λ'=−0.13	**46.8**	**53.9**	**57.9**	**62.3**	**63.8**	**32.6**	**37.4**	**44.4**	**47.1**	**48.4**	**39.4**	**45.9**	**50.9**	**58.2**	**59.3**

注：黑体数值表示最优值，DC 包含漂移补偿和分布校准，LP-FBR 表示基于首项偏差削减的线性探测，OFT-IFBR 表示基于逆首项偏差削减的整体微调，λ 和 λ'表示 2 步微调的超参数，“√”表示使用该策略。

首先利用超参数 λ=0.7 时首项偏差削减的线性探测策略，再利用不同的逆首项偏差削减的整体微调策略观察基类和新类的检测性能。从图 7 所示演示结果可知，逆首项偏差削减在 λ'=−0.13 时效果最好。

同时可知，在 λ' >0 时，逆首项偏差削减转化为首项偏差削减，引起过高置信度导致的过参数化问题，进而使基类和新类的性能下降。

3）蒸馏学习的消融研究

图 8 给出了不同知识蒸馏方法在 PASCAL VOC 数据集的检测结果，其中兰德指数（Rand index, RI）值越大表示特征聚类质量越好。从图 8 可知，与跨层知识蒸馏（cross-level knowledge distillation, CLD）方法相比，同层知识蒸馏（same-level knowledge distillation, SLD）方法具有更好的聚类性能；与同级蒸馏方法相比，1 对多知识蒸馏（one-to-all knowledge distillation, OAD）方法获得了更高的兰德指数值，表明学生模型的浅层模块可以学习新类样本的判别特征。

最后，本文提出的跨模块知识蒸馏方法 CMD 在每个特征提取模块中都获得最高的兰德指数值，这表明本文提出的 CMD 可以提取新类样本的更多判

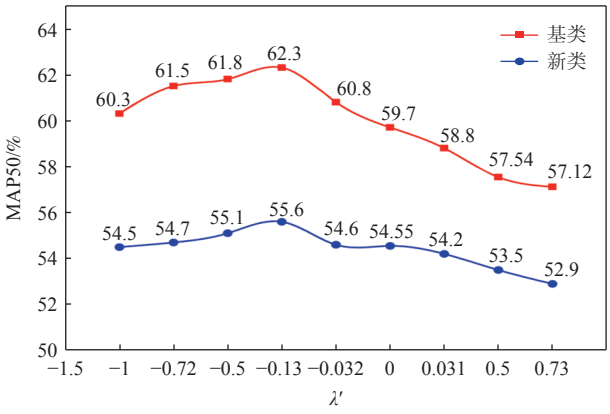


Fig. 7 Impact of hyperparameter λ' on detection performance in inverse first bias reduction

图 7 逆首项偏差削减中超参数 λ'对于检测性能的影响

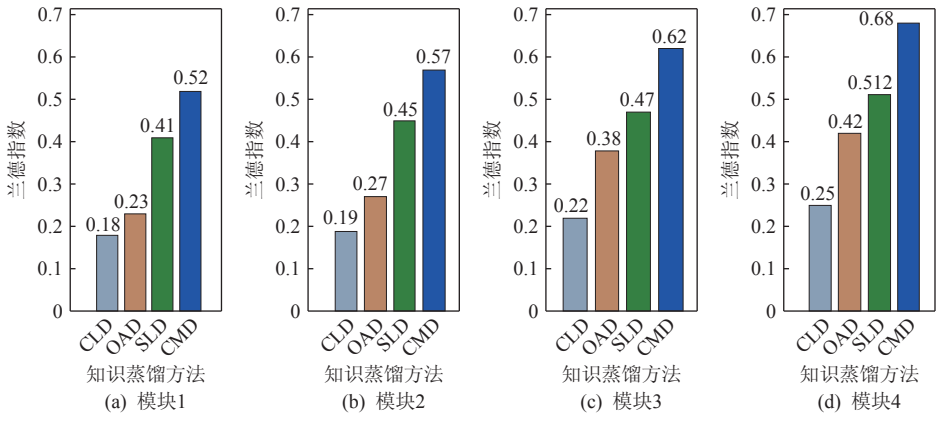
别特征，并有效缓解从教师模型到学生模型的有偏的新类知识转移问题。

图 9 给出不同知识蒸馏方法的示意图，其中图 9(a)中的跨层知识蒸馏方法映射学生模型的浅层特征，并使用教师模型的深层特征计算蒸馏损失；图 9(b)中的同层知识蒸馏方法，使用学生模型和教师模型的同层特征计算蒸馏损失；图 9(c)中的 1 对多知识蒸馏方法利用教师模型映射学生模型的浅层特征，并计算教师模型最后 1 层的蒸馏损失；图 9(d)中的跨模块知识蒸馏方法使用教师模型的深层模块映射学生模型的浅层特征，并使用教师模型的深层特征计算蒸馏损失。

表 6 给出了不同知识蒸馏方法和不同训练迭代间隔 ϖ 对学生模型检测性能的影响。从表 6 可知，与跨层知识蒸馏方法相比，在 5 个新类样本的条件下，同层知识蒸馏方法的性能分别提高了 2.2 个百分点、1.1 个百分点、1.0 个百分点；与同层知识蒸馏方法相比，在 1 个新类样本的条件下，1 对多知识蒸馏方法的性能分别提高了 0.1 个百分点、0.1 个百分点、0.2 个百分点；与 1 对多知识蒸馏方法相比，在 3 个新类样本的条件下，CMD 的性能分别达到 59.9%，45.8%，52.4%；上述实验结果表明，与 CMD 相比，其他 3 种知识蒸馏方法可提取更多的判别特征。此外，在 1 个新类样本的条件下，当先前学生模型添加至原始跨模块知识蒸馏方法后，模型的性能分别提高了 0.6 个百分点、0.7 个百分点、0.9 个百分点；当训练迭代间隔 ϖ=3 时，学生模型的性能达到最佳。

4)超参数分析

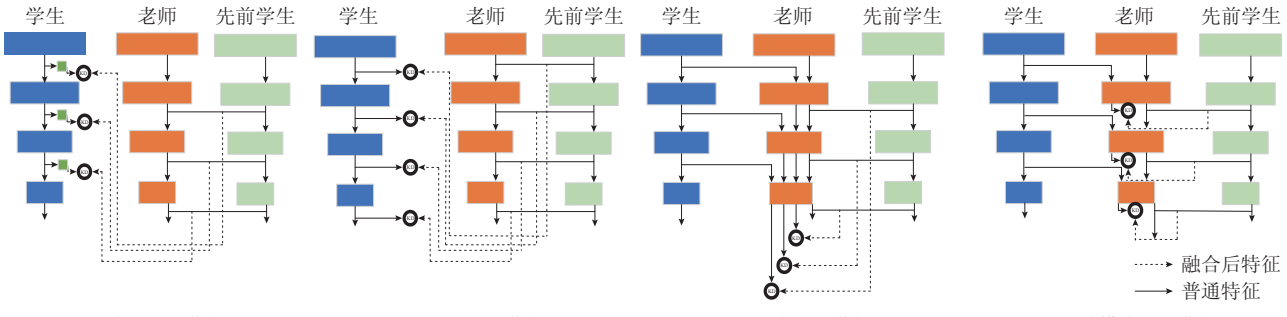
表 7 给出了整体微调过程中，损失函数中的超参



(a) 模块1 (b) 模块2 (c) 模块3 (d) 模块4

Fig. 8 Detection performance of different knowledge distillation methods

图 8 不同知识蒸馏方法的检测性能



(a) 跨层知识蒸馏 (b) 同层知识蒸馏 (c) 1对多知识蒸馏 (d) 跨模块知识蒸馏

Fig. 9 Different distillation methods used for aligning features between student model and teacher model

图 9 用于对齐学生模型和教师模型特征的不同蒸馏方法

Table 6 Ablation Experiments on PASCAL VOC Dataset in Terms of MAP50

表 6 在 PASCAL VOC 数据集的以 MAP50 为评估指标的消融实验 %

索引	DC	LP-FBR	OFT-IFBR	蒸馏方法	第 1 组					第 2 组					第 3 组				
					1 个	2 个	3 个	5 个	10 个	1 个	2 个	3 个	5 个	10 个	1 个	2 个	3 个	5 个	10 个
1	√	λ=0.7	λ'=−0.13	CLD[27]	46.8	53.9	57.9	62.3	63.8	32.6	37.4	44.4	47.1	48.4	39.4	45.9	50.9	58.2	59.3
2	√	λ=0.7	λ'=−0.13	SLD	47.5	54.7	59.8	64.5	65.9	33.7	38.6	45.6	48.2	51.3	40.6	47.1	52.1	59.2	60.7
3	√	λ=0.7	λ'=−0.13	OAD	47.6	54.8	60.3	65.9	66.7	33.8	38.7	44.9	48.5	51.7	40.8	47.2	52.9	59.4	61.2
4	√	λ=0.7	λ'=−0.13	CMD（OSM）	47.8	55.0	59.9	64.9	66.1	33.9	38.9	45.8	48.6	51.7	40.9	47.5	52.4	59.6	60.9
5	√	λ=0.7	λ'=−0.13	CMD（ϖ=0）	48.4	55.9	61.6	67.0	67.9	34.6	39.8	47.0	49.7	52.8	41.8	48.5	54.0	60.6	62.4
6	√	λ=0.7	λ'=−0.13	CMD（ϖ=1）	48.8	57.0	63.6	67.4	68.2	34.8	42.0	48.1	50.4	53.7	42.2	49.4	54.9	61.5	63.9
7	√	λ=0.7	λ'=−0.13	CMD（ϖ=3）	**50.1**	**58.2**	**64.4**	**68.5**	**68.6**	**35.2**	**43.1**	**49.0**	**51.1**	**54.4**	**43.3**	**51.2**	**55.8**	**63.0**	**64.7**
8	√	λ=0.7	λ'=−0.13	CMD（ϖ=5）	48.3	55.8	61.3	66.9	67.7	34.4	39.7	46.9	49.5	52.7	41.6	48.2	53.9	60.4	62.2
9	√	λ=0.7	λ'=−0.13	CMD（ϖ=7）	48.0	54.9	60.7	66.5	66.9	34.0	39.3	46.2	49.1	52.1	41.2	48.0	53.3	60.1	61.5
10	√	λ=0.7	λ'=−0.13	CMD（ϖ=10）	47.7	54.4	60.2	66.2	66.3	33.5	38.4	45.7	48.5	51.6	40.4	47.4	52.1	59.4	61.2

注：黑体数值表示最优值，DC 包含漂移补偿和分布校准，LP-FBR 表示基于首项偏差削减的线性探测，OFT-IFBR 表示基于逆首项偏差削减的整体微调，OSM 表示不使用旧学生模型，“√”表示使用该策略。

Table 7 Ablation Experiments of Hyperparamet η on PASCAL VOC Dataset in Terms of MAP50

表 7 PASCAL VOC 数据集的以 MAP50 为评估指标的超参数 η 消融实验 %

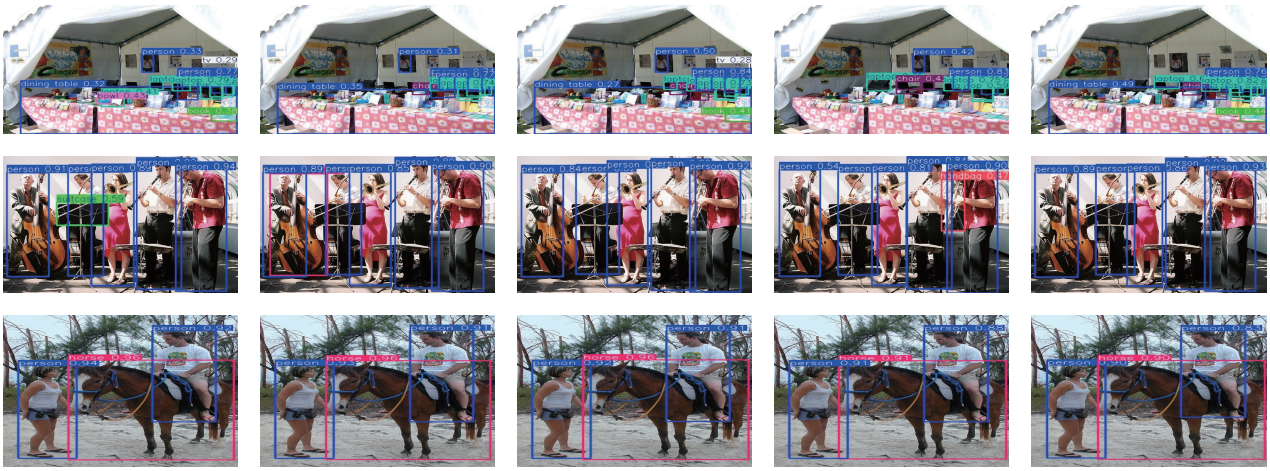
η	第 1 组					第 2 组					第 3 组				
	1 个	2 个	3 个	5 个	10 个	1 个	2 个	3 个	5 个	10 个	1 个	2 个	3 个	5 个	10 个
0.1	49.8	57.9	64.1	67.8	68.2	34.9	42.8	48.8	50.9	54.1	42.7	50.8	55.3	62.6	64.1
0.3	50.4	58.3	64.5	68.3	68.7	35.4	43.2	49.1	51.3	54.5	43.1	51.5	55.6	63.2	64.5
0.5	50.6	58.6	64.9	68.8	68.9	35.5	43.6	49.4	51.7	54.8	43.5	51.8	56.1	63.5	64.9
0.7	**51.1**	**58.9**	**65.2**	**69.3**	**69.1**	**35.7**	**44.0**	**49.8**	**52.2**	**55.3**	**43.9**	**52.1**	**56.7**	**63.8**	**65.4**
1.0	50.1	58.2	64.4	68.5	68.6	35.2	43.1	49.0	51.1	54.4	43.3	51.2	55.8	63.0	64.7

注：黑体数值表示最优值。

数 η 对模型性能的影响。由表 7 可知：当 η 从 0.1 逐渐增加至 0.7 时，模型性能不断提高，这表明在此过程中教师模型引导学生模型在浅层模块中学习深层语义信息，这有助于提取新类更多的判别特征；当 η 从 0.7 增加至 1.0 时，模型性能逐渐降低，这表明过大的蒸馏损失会导致教师模型向学生模型传递有偏的新类特征。因此，实验中选择超参数 η=0.7 时模型性能最佳。

3.5 可视化分析

图 10 给出 MS COCO 数据集上 5 组对比实验检



本文方法 VFA[35] Meta Faster-RCNN[38] MPF-Net[42] FPD[39]

Fig. 10 Detection performance of different methods on MS COCO dataset

图 10 不同方法在 MS COCO 数据集的检测性能

测结果,其中测试的3张图片中有基类和新类,基类有人、餐桌、笔记本电脑,新类有碗、书、电视机、手提箱、马。由图10可知,与其他方法对比,本文方法的新类样本错检和漏检情况有所减少,同时目标预测准确度有所提高;此外,本文方法中,即使基类和新类的特征差异较大时,基类和新类间的预测准确度和定位准确度基本接近,这说明本文方法有效缓解了特征差异过大导致知识迁移困难问题。

4 结 论

为解决少样本目标检测中无效的鲁棒性假设问题,本文提出了基于跨模块知识蒸馏的3阶段微调的少目标检测方法。首项偏差削减策略,有效缓解了2步微调的线性探测过程中,因与基类语义相似性较低的新类样本生成有偏特征,导致模型学习有偏新类知识;逆首项偏差削减策略,有效缓解了2步微调的整体微调过程中因与基类具有高度相似的新类引起的特征提取网络的过拟合问题。在公开数据集上大量的对比实验和消融实验验证了所提方法的优越性,本文所提的少样本目标检测方法有效提高了检测性能。

作者贡献声明:王双丞提出研究思路、编写代码,并撰写论文;朱松豪负责设计和完成对比实验,并修订论文。

参 考 文 献

[1] Wang Xin, Huang E, Darrell T, et al. Frustratingly simple few-shot object detection[C] //Proc of the 37th Int Conf on Machine Learning. Berlin: Springer, 2020: 9919−9928

[2] Kumar A, Raghunathan A, Jones R, et al. Fine-tuning can distort pretrained features and underperform out-of-distribution[C] //Proc of the 10th Int Conf on Learning Representations. Cambridge, MA: MIT, 2022: 1−42

[3] Wang Jiabao, Chen Yuming, Zheng Zhaohui, et al. CrossKD: Cross-head knowledge distillation for dense object detection[J]. Computing Research Repository, 2023, 31(1): 1−15

[4] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C] //Proc of the 34th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 779−788

[5] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C] //Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 213−229

[6] Chen Honghui, Zheng Jianming, Cai Fei, et al. Modeling of few-shot relation extraction based on adaptive self-training[J]. Journal of Computer Research and Development, 2023, 60(7): 1581−1591 (in Chinese)
(陈洪辉, 郑建明, 蔡飞, 等. 基于适应性自训练的少样本关系抽取建模[J]. 计算机研究与发展, 2023, 60(7): 1581−1591)

[7] Yang Jieyi, Dong Yihong, Qian Jiangbo. Research progress of few-shot learning methods based on graph neural networks[J]. Journal of Computer Research and Development, 2024, 61(4): 856−876 (in Chinese)
(杨洁祎, 董一鸿, 钱江波. 基于图神经网络的小样本学习方法研究进展[J]. 计算机研究与发展, 2024, 61(4): 856−876)

[8] Liu Binghao, Ding Yao, Jiao Jianbin, et al. Anti-aliasing semantic reconstruction for few-shot semantic segmentation[C] //Proc of the 39th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 9747−9756

[9] Xu Jian, Liu Bo, Xiao Yanshan. A variational inference method for few-shot learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 33(1): 269−282

[10] Shao Shuai, Xing Lei, Wang Yanjiang, et al. Attention-based multi-view feature collaboration for decoupled few-shot learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 33(5): 2357−2369

[11] Tang Yingbo, Cao Zhiqiang, Yang Yuequan, et al. Semi-supervised few-shot object detection via adaptive pseudo labeling[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(4): 2151−2165

[12] Ma Jiawei, Niu Yulei, Xu Jincheng, et al. Digeo: Discriminative geometry-aware learning for generalized few-shot object detection[C] // Proc of the 41st IEEE Conf on Computer Vision and Pattern Recognition. New York: IEEE, 2023: 3208−3218

[13] Yang Shuo, Liu Lu, Xu Min. Free lunch for few-shot learning: Distribution calibration[C] //Proc of the 9th Int Conf on Learning Representations. Berlin: Springer, 2021: 1−13

[14] Zhang Songyang, Li Zeming, Yan Shipeng, et al. Distribution alignment: A unified framework for long-tail visual recognition[C] // Proc of the 39th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 2361−2370

[15] Song Hao, Diethe T, Kull M, et al. Flach. Distribution calibration for regression[C] //Proc of the 36th Int Conf on Machine Learning. Berlin: Springer, 2019: 5897−5906

[16] Shen Zhiqiang, Liu Zechun, Qin Jie, et al. S2-BNN: Bridging the gap between self-supervised real and 1-bit neural networks via guided distribution calibration[C] //Proc of the 39th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 2165−2174

[17] Li Bohao, Liu Chang, Shi Mengnan, et al. Proposal distribution calibration for few-shot object detection[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(1): 1911−1918

[18] Zhao Zhiyuan, Liu Qingjie, Wang Yunhong. Exploring effective knowledge transfer for few-shot object detection[C] //Proc of the 30th ACM Int Conf on Multimedia. New York: ACM, 2022: 6831−6839

[19] Wu Jiashan, Lang Chunbo, Cheng, Gong, et al. Retentive

compensation and personality filtering for few-shot remote sensing object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(7): 5805−5817

[20] Wu Jiaying, Hu Jinglu. Redefining prior feature space via fine-tuning a triplet network for few-shot learning[J]. IET Computer Vision, 2022, 16(6): 514−524

[21] Tian Yonglong, Wang Yue, Dilip Krishnan, et al. Rethinking few-shot image classification: A good embedding is all you need?[C] //Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 266−282

[22] Mangla P, Singh M, Sinha A, et al. Charting the right manifold: Manifold mixup for few-shot learning[C] //Proc of the 15th IEEE Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2020: 2218−2227

[23] Rajasegaran J, Khan S, Hayat M, et al. Self-supervised knowledge distillation for few-shot learning[C] //Proc of the 32nd British Machine Vision Conf. Piscataway, NJ: IEEE, 2021: 1−14

[24] Tao Ran, Zhang Han, Zheng Yutong, et al. Powering finetuning in few-shot learning: Domain-agnostic bias reduction with selected sampling[C] //Proc of the 36th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2022: 8467−8475

[25] Wang Heng, Yue Tan, Ye Xiang, et al. Revisit finetuning strategy for few-shot learning to transfer the emdeddings[C] //Proc of the 18th Int Conf on Learning Representation. Berlin: Springer, 2023: 1−11

[26] Li Jiangmeng, Zhang Yanan, Qiang Wenwen, et al. Disentangle and remerge: Interventional knowledge distillation for few-shot object detection from a conditional causal perspective[C] //Proc of the 37th AAAI Conf on Artificial Intelligence (AAAI'2023). Palo Alto, CA: AAAI, 2023: 1323−1333

[27] Zheng Hao, Wang Runqi, Liu Jianzhuang, et al. Cross-level distillation and feature denoising for cross-domain few-shot classification[C] //Proc of the 18th Int Conf on Learning Representations. Berlin: Springer, 2023: 1−14

[28] Yu Ying, Wei Wei, Tang Hong, et al. Multi-stage training with multi-level knowledge self-distillation for fine-grained image recognition[J]. Journal of Computer Research and Development, 2023, 60(8): 1834−1845 (in Chinese)
(余鹰, 危伟, 汤洪, 等. 多层次知识自蒸馏联合多步骤训练的细粒度图像识别[J]. 计算机研究与发展, 2023, 60(8): 1834−1845)

[29] Firth D. Bias reduction of maximum likelihood estimates[J]. Biometrika, 1993, 80(1): 27−38

[30] Ghaffari S, Saleh E, Forsyth D, et al. On the importance of firth bias reduction in few-shot classification[C] //Proc of the 17th Int Conf on Learning Representations. Berlin: Springer, 2022: 1−29

[31] Müller R, Kornblith S, Hinton G. When does label smoothing help?[C] //Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 4696−4705

[32] Hein M, Andriushchenko M, Bitterwolf J. Why relu networks yield high confidence predictions far away from the training data and how to mitigate the problem[C] //Proc of the 37th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 41−50

[33] Sun Bo, Li Banghuai, Cai Shengcai, et al. FSCE: Few-shot object detection via contrastive proposal encoding[C] //Proc of the 39th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 7352−7362

[34] Wu Jiaxi, Liu Songtao, Huang Di, et al. Multi-scale positive sample refinement for few-shot object detection[C] //Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 456−472

[35] Han Jiaming, Ren Yuqiang, Ding Jian, et al. Few-shot object detection via variational feature Aggregation[C] //Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 755−763

[36] Hu Hanzhe, Bai Shuai, Li Aoxue, et al. Dense relation distillation with context-aware aggregation for few-shot object detection[C] //Proc of the 39th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 10185−10194

[37] Zhang Shan, Wang Lei, Murray N, et al. Kernelized few-shot object detection with efficient integral aggregation[C] //Proc of the 40th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 19207−19216

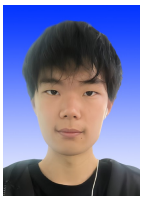
[38] Han Guangxing, Huang Shiyuan, Ma Jiawei, et al. Meta Faster R-CNN: Towards accurate few-shot object detection with attentive feature alignment[C] //Proc of the 36th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2022: 780−789

[39] Wang Zichen, Yang Bo, Yue Haonan, et al. Fine-grained prototypes distillation for few-shot object detection[C] //Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 5859−5866

[40] Xiao Yang, 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

[41] Zhang Weilin, Wang Yuxiong. Hallucination improves few-shot object detection[C] //Proc of the 39th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 13008−13017

[42] Han Chen, Wang Qi, Xie Kailin, et al. MPF-Net: Multi-projection filtering network for few-shot object detection[J]. Applied Intelligence, 2024, 54(17): 7777−7792



Wang Shuangcheng, born in 2002. Master candidate. His main research interests include pattern recognition and machine learning.
王双丞, 2002年生。硕士研究生。主要研究方向为模式识别、机器学习。



Zhu Songhao, born in 1973. PhD. His main research interests include image processing, pattern recognition, and artificial intelligence.
朱松豪, 1973年生。博士。主要研究方向为图像处理、模式识别、人工智能。