计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(11):270−278 [doi: 10.15888/j.cnki.csa.009988] [CSTR: 32024.14.csa.009988]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

空间-通道注意力协同的精炼特征知识蒸馏①

叶超越, 钟良琪, 闫胜业

(南京信息工程大学 自动化学院, 南京 210044)
通信作者: 闫胜业, E-mail: shengye.yan@gmail.com

摘　要: 知识蒸馏通过传递教师模型知识提升学生模型性能. 然而对于轻量化的学生模型而言, 全盘吸收教师特征图内的隐含知识是困难的, 为此本文提出一种基于空间-通道注意力精炼的协同蒸馏方法 (SCAR-KD), 通过从原始特征图中提炼关键判别信息, 缓解学生与教师的语义分歧. 具体而言, 本文采用了一种多尺度空间-通道注意力模块 (SCSA), 从师生特征图的通道和空间维度中精炼出具有判别性的注意力增强特征进行蒸馏, 同时解耦出空间注意力图加权给原始特征图进行动态蒸馏. 该方法实现了双重知识迁移. 实验结果表明, YOLOv8n-SCAR-KD 相较于基线 YOLOv8n 在 VOC 和 VisDrone 数据集上 mAP@0.5:0.95 分别从 64.1% 提升至 65.3%, 从 20.7% 提升至 21.6%, 超过了现有的主流蒸馏方法, 验证了方法的有效性.

关键词: 知识蒸馏; 目标检测; 特征模仿; 注意力机制; 语义分歧

引用格式: 叶超越,钟良琪,闫胜业.空间-通道注意力协同的精炼特征知识蒸馏.计算机系统应用,2025,34(11):270–278. http://www.c-s-a.org.cn/1003-3254/9988.html

Refined Feature Knowledge Distillation Based on Spatial-channel Collaborative Attention

YE Chao-Yue, ZHONG Liang-Qi, YAN Sheng-Ye

(School of Automation, Nanjing University of Information Science & Technology, Nanjing 210044, China)

Abstract: Knowledge distillation enhances student model performance by transferring knowledge from the teacher model. However, for lightweight student models, it is challenging to fully absorb the implicit knowledge contained in the teacher feature maps. To address this, this study proposes a spatial-channel attention refinement-based collaborative distillation (SCAR-KD) method, which extracts key discriminative information from the original feature maps to mitigate the semantic discrepancy between the student and teacher. Specifically, we adopt a multi-scale spatial-channel attention (SCSA) module to refine discriminative attention-enhanced features from the teacher feature maps along both the channel and spatial dimensions, while decoupling the spatial attention map to weight the original feature maps for dynamic distillation. This approach achieves dual knowledge transfer. Experimental results demonstrate that the mAP@0.5:0.95 of YOLOv8n-SCAR-KD model improved from 64.1% to 65.3% on the VOC dataset and from 20.7% to 21.6% on the VisDrone dataset, outperforming current mainstream distillation methods and validating the effectiveness of the proposed method.

Key words: knowledge distillation (KD); object detection; feature imitation; attention mechanism; semantic discrepancy

目标检测是计算机视觉领域的核心任务, 在实时场景中兼顾高精度与高效性仍具挑战. 现有检测方法计算开销较高, 限制了其在资源受限设备上的应用. YOLOv8[1]作为高效目标检测框架, 虽能以小型模型实现良好性能, 但与大模型仍存在较大差距. 知识蒸馏[2]是一种模型压缩技术, 通过教师模型指导学生模型学

① 收稿时间: 2025-04-02; 修改时间: 2025-04-27, 2025-05-12; 采用时间: 2025-05-19; csa 在线出版时间: 2025-09-30
CNKI 网络首发时间: 2025-10-09

习, 以提升其检测精度. 本文探讨如何利用知识蒸馏技术优化轻量级 YOLOv8 的检测性能.

本文提出了一种基于空间-通道注意力精炼的协同蒸馏方法 (SCAR-KD), 构建于教师-学生蒸馏框架之上.

近年来, 知识蒸馏技术通过迁移教师模型的深层知识来提升学生模型性能, 已成为模型压缩领域的重要研究方向. 随着研究的深入, 对齐特征图挖掘教师模型深层丰富的语义信息成为研究主流. 然而, 这种特征对齐方法面临一个关键挑战: 轻量化的学生模型由于结构容量有限, 难以充分吸收教师模型复杂的多层次表征, 导致显著的“知识代沟”问题. 针对这一挑战, 现有研究主要采用两种解决思路. 第 1 种是通过引入助教模型[3,4]来桥接教师和学生模型. 虽然这种方法在一定程度上缓解了知识迁移的难度, 但也带来了一定的局限性: 助教模型的设计需要大量经验性调参. 第 2 种思路是设计复杂的损失衰减曲线, 之前的工作[5]表明训练初期教师提供强监督, 随后逐步减弱, 使学生独立学习可以帮助学生过渡. 但该方法依赖手工设计权重函数.

本文考虑 SCSA 注意力能够挖掘多语义信息的协同潜力来进行特征引导和缓解语义差异的特性, 我们设计了一种基于空间-通道注意力 (SCSA) 的特征蒸馏框架. 具体而言, 提出一种双路径知识迁移机制: 首先, 通过通道-空间双维度特征精炼模块提取教师特征图中的判别性特征, 抑制背景噪声等无关信息, 生成精炼后的特征图. 其次利用空间注意力图对师生原始特征图进行自适应加权, 突出关键区域特征. 在此基础上, 设计双重知识蒸馏路径: (1) 教师精炼特征与学生精炼特征之间的高阶语义对齐; (2) 学生增强空间特征与教师增强空间特征之间的结构化特征匹配. 这种双路径设计既保留了教师模型的不同阶段的空间信息, 又强化了学生对关键特征区域的学习能力.

本文基于文献[5]的层级加权特征蒸馏 (LWFI) 工作进行扩展改进, 总结而言, 我们的贡献主要是以下 3 方面: (1) 发现了之前方法的不足之处: 对于轻量化学生模型学习能力有限情况下, 全盘学习教师的特征图是困难的. (2) 提出了一种改进的知识蒸馏方法——基于空间-通道注意力精炼的协同蒸馏方法, 用于缓解教师和学生之间的知识代沟. (3) 把该方法与当前主流的特征蒸馏方法比较, 在多个公共数据集上验证了其有效性.

1 相关工作

1.1 目标检测中特征知识蒸馏

知识蒸馏最早由 Hinton 等人[2]在 2015 年提出, 最初应用于分类任务, 通过使用教师网络的输出作为软标签来辅助学生模型的学习过程. 随后, Zagoruyko 等人[6]提出了注意力转移, 通过将教师网络的注意力图转移到学生网络, 帮助学生网络学习更好的注意力机制. 为了解决目标检测中的前景-背景类别不平衡问题, Yang 等人[7]提出了 FGD, 结合了局部 (focal) 和全局 (global) 知识蒸馏. 局部知识蒸馏关注于模型的局部特征, 而全局知识蒸馏则关注于模型的整体结构. 之后, Yang 等人[8]提出了 MGD 方法, 学生网络通过随机模式遮蔽特征图, 并通过生成网络近似教师网络的特征图, 从而提高模型的泛化能力. Shu 等人[9]提出了通道蒸馏 CWD 方法, 该方法通过归一化每个通道的特征图, 获得软概率图, 并通过最小化两个网络之间的 KL 散度, 使蒸馏过程更加关注每个通道的重要区域. RKD[10]通过提取特征图间基于距离和角度的相关性关系进行蒸馏. 随后, Yang 等人[11]提出了 BCKD 方法, 解决了密集目标检测中的类别不平衡问题, 提出了二分类蒸馏损失, 改进了传统的分类蒸馏方法. CrossKD[12]将学生检测头的中间特征传递给教师的检测头, 然后强制这些交叉头预测模仿教师的预测.

1.2 注意力机制

该机制使得模型能够更精准地从图像中提取局部上下文信息. 部分方法如 (SE-NET[13]和 SGE-NET[14]) 在通道间层面计算注意力, 而自注意力机制则考虑输入像素的成对相似性. BAM[15]同步计算通道与空间注意力, CBAM[16]在通道注意力后执行空间注意力操作, FcaNet[17]提出频域通道注意力机制. EMA[18]通过通道分组实现高效多尺度注意力. 近期考虑到通道和空间注意之间的协同作用尚未得到充分挖掘, 试图在多个语义层面揭示空间和通道注意之间的协同关系, 提出了一个空间和通道协同注意模块 SCSA[19].

在知识蒸馏领域, 多项研究已引入注意力机制. 文献[20]利用自注意力软化 logits 以简化学习过程. 文献[21]对特征施加位置自注意力进行蒸馏. SAKD[22]以学生特征为查询项、教师特征为键值项计算稀疏注意力, 并通过随机失活实现注意力优化. 然而. 现有方法多侧重于单一维度的注意力建模, 未能充分挖掘空间

与通道注意力之间的协同关系. 故本文融合空间-通道的协同注意力机制进行进一步探索.

2 方法描述

2.1 回顾层级加权特征蒸馏

在之前提出的增强式层级加权特征蒸馏方法中, 特征蒸馏涉及从网络的骨干、颈部和头部提取多尺度特征图, 并对这些特征图应用软加权, 以自动适应各部分的表征能力. 具体而言, 采用 Softmax 操作为不同的特征层分配权重, 从而动态调整每一层的重要性. 然后, 利用 Kullback-Leibler (KL) 散度来衡量教师模型与学生模型在特征层面的差异, 从而计算蒸馏损失, KL 是通常用于衡量数据之间的分布差异, 我们的目标是最小化教师和学生之间的 KL 差异. 整体的层级加权特征结构图如图 1 所示. 其中骨干、颈部、头部提供用于特征蒸馏的特征图信息, 特征蒸馏损失可以表示如下.

$$L_{\text{fea}} = \alpha KL\left(f_{\text{b}}^{\text{t}} \| f_{\text{b}}^{\text{s}}\right) + \beta KL\left(f_{\text{n}}^{\text{t}} \| f_{n}^{\text{s}}\right) + \gamma KL\left(f_{\text{h}}^{\text{t}} \| f_{\text{h}}^{\text{s}}\right) \quad (1)$$

其中, f_{b}^{t}和f_{b}^{s}分别表示教师模型和学生模型在骨干网络中的特征图. 同样, f_{n}^{t}、f_{n}^{s}和f_{h}^{t}、f_{h}^{s}分别代表颈部网络和检测头网络的特征图. 参数α、β、γ表示在上述 3 个网络位置上的特征蒸馏损失的权重. 在后续的实验中, 我们比较了不同参数取值对性能的影响, 并最终采用 Softmax 方法自动确定这些参数的值, 如下所示:

$$(\alpha,\beta,\gamma) = S\left(\frac{KL(f_{\text{b}}^{\text{t}} \| f_{\text{b}}^{\text{s}})}{n_{\text{b}}}, \frac{KL(f_{\text{n}}^{\text{t}} \| f_{\text{n}}^{\text{s}})}{n_{\text{n}}}, \frac{KL(f_{\text{h}}^{\text{t}} \| f_{\text{h}}^{\text{s}})}{n_{\text{h}}}\right) \quad (2)$$

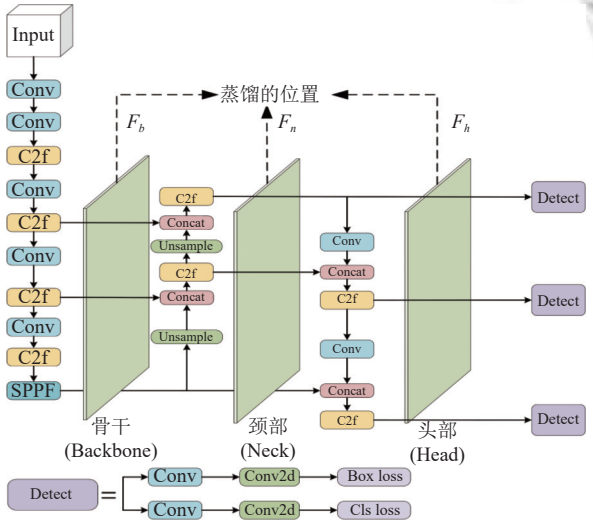


图 1 YOLOv8 的层级加权特征蒸馏

其中, n_{b}、n_{n}、n_{h}分别表示骨干、颈部和头部网络的分支数, S 代表 Softmax 操作. 该方法使得不同层次的特征对蒸馏损失的贡献动态变化, 进一步提高蒸馏效率.

2.2 SCSA 空间-通道注意力模块

为进一步增强教师的特征语义信息, 本文采用的空间-通道注意力 (SCSA) 模块由空间多语义注意力 (SMSA) 与通道级自注意力 (PCSA) 两部分组成, 能够同时建模空间分布特征与通道间相关性, 其整体结构如图 2 所示.

2.2.1 空间多语义注意力模块

神经网络架构中, 分解技术能够有效降低参数量和计算开销, 该模块中对输入特征图$X \in \mathbb{R}^{B\times C\times H\times W}$进行高度和宽度方向的分解, 具体而言对高度和宽度方向分别进行全局平均池化, 形成两个一维序列$X_H \in \mathbb{R}^{B\times C\times W}$, $X_W \in \mathbb{R}^{B\times C\times H}$, 为了捕捉不同的空间分布和上下文关系, 将特征划分为 n 组等大小, 独立的子特征X_H^i和X_W^i, 每个子特征通道数为$\frac{C}{n}$, 特征分解的具体方式如下:

$$X_H^i = X_H\left[:,(i-1)\times\frac{C}{n}:i\times\frac{C}{n},:\right] \quad (3)$$

$$X_W^i = X_W\left[:,(i-1)\times\frac{C}{n}:i\times\frac{C}{n},:\right] \quad (4)$$

其中, X^i代表第i个子特征, $i\in[1,n]$, 将特征图划分为不同子特征后, 为了减少语义空白, 我们通过并行施加核尺寸为 (3, 5, 7, 9) 的深度可分离一维卷积, 针对 H/W 维度分解导致的感受野受限问题, 采用轻量化共享卷积进行特征对齐, 多语义空间信息提取的实现如下:

$$\tilde{X}_H^i = DWConv1d_{k_i}^{\frac{C}{n}\to\frac{C}{n}}(X_H^i) \quad (5)$$

$$\tilde{X}_W^i = DWConv1d_{k_i}^{\frac{C}{n}\to\frac{C}{n}}(X_W^i) \quad (6)$$

在完成独立子特征分解及异构语义空间信息提取后, 需构建空间注意力图. 我们拼接不同语义子特征, 并采用分组归一化 (GN) 进行标准化处理 (组数与子特征数一致). 相比传统批量归一化 (BN), GN 能更有效区分子特征间的语义差异: 其独立归一化机制避免了批统计噪声干扰, 最终通过 Sigmoid 函数生成空间注意力权重, 实现关键区域激活与非重要区域抑制. 特征输出计算式如下:

$$Attn_H = \sigma\left(GN_H^n\left(Concat\left(\tilde{X}_H^1,\tilde{X}_H^2,\cdots,\tilde{X}_H^n\right)\right)\right) \quad (7)$$

$$Attn_W = \sigma\left(GN_W^n\left(Concat\left(\tilde{X}_W^1, \tilde{X}_W^2, \cdots, \tilde{X}_W^n\right)\right)\right) \quad (8)$$

$$SMSA(X) = X_s = Attn_H \times Attn_W \times X \quad (9)$$

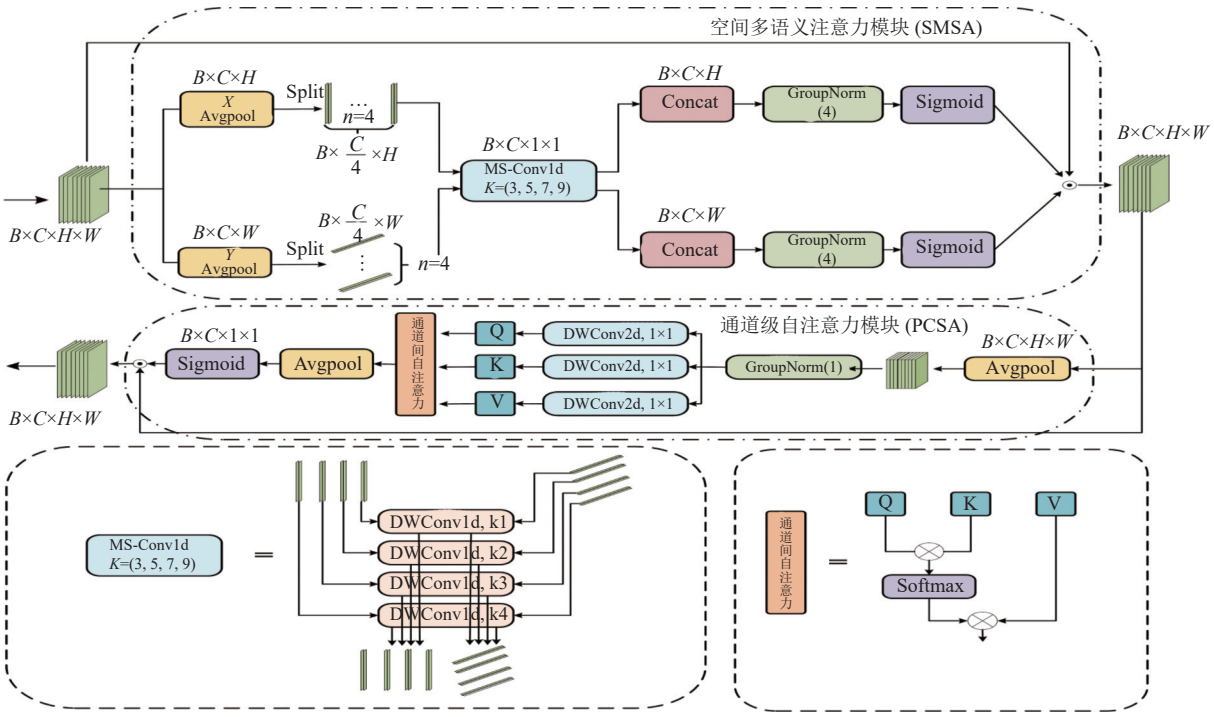


图 2 SCSA 空间-通道注意力模块示意图

2.2.2 通道级自注意力模块

传统基于卷积的通道注意力方法存在固有局限: 其卷积运算难以直观建模通道间相似性. 受 ViT[23]中多头自注意力 (MHSA) 在 token 相似性建模的启发, 该模块将单头自注意力 (SHSA) 与 SMSA 提供的调制空间先验相结合, 以更高效计算通道间相似度. 为保持 SMSA 提取的多语义空间信息并降低 SHSA 计算开销, 基于平均池化的渐进压缩机制作为协同引导. 相较于传统卷积方法, PCSA 模块具有更强的输入感知能力, 并能有效利用 SMSA 空间先验深化表征学习. 具体实现如下.

$$X_p = Pool_{(7,7)}^{(H,W)\to(H',W')}(X_s) \quad (10)$$

$$F_{\text{proj}} = DWConv1d_{(1,1)}^{C\to C} \quad (11)$$

$$Q = F_{\text{proj}}^{Q}\left(X_p\right),\ K = F_{\text{proj}}^{K}\left(X_p\right),\ V = F_{\text{proj}}^{V}\left(X_p\right) \quad (12)$$

$$X_{\text{attn}} = Attn(Q, K, V) \quad (13)$$

$$PCSA\left(X_s\right) = X_s \times \sigma\left(Pool_{(H',W')}^{(H',W')\to(1,1)}\left(X_{\text{attn}}\right)\right) \quad (14)$$

其中, $Pool_{(k,k)}^{(H,W)\to(H',W')}(\cdot)$代表了核尺寸 (k, k) 的池化操作, 将特征图分辨率从 (H, W) 下采样到(H', W'). $F_{\text{proj}}(\cdot)$ 代表了线性映射层, 用于生成 (Q, K, V).

2.3 双重路径精炼蒸馏

为实现更有效的知识迁移, 本文提出了双重路径精炼蒸馏 (SCAR-KD), 其整体框架如图 3 所示. 该方法包含两个分支: 动态加权蒸馏分支和精炼特征图蒸馏分支.

对于特征图动态加权蒸馏部分, 其中教师网络的注意力图是教师原始特征图通过空间多语义模块得到的$Attn_H \times Attn_W$, 作为知识迁移的载体, 在特征图蒸馏中需要将其关注的重点区域通过加权方式传递给学生网络, 这一设计灵感来源于现实教学中教师为学生划重点的教学方法, 帮助学生知道哪里是关键区域, 对于关键位置加强学习, 我们结合了之前层级加权蒸馏的工作, 对于师生特征图的蒸馏同样采用了动态感知损失加权的策略.

$$\overline{f_s} = f_s \odot (Attn_H \times Attn_W) \quad (15)$$

$$\overline{f_t} = f_t \odot (Attn_H \times Attn_W) \quad (16)$$

$$L_{\text{fea}} = \alpha KL\left(\overline{f_b^t} \| \overline{f_b^s}\right) + \beta KL\left(\overline{f_n^t} \| \overline{f_n^s}\right) + \gamma KL\left(\overline{f_h^t} \| \overline{f_h^s}\right) \quad (17)$$

对于精炼特征图蒸馏分支, F_t'和F_s'表示经 SCSA 模块处理后的精炼特征图, SMSA 模块通过门控机制生成的空间注意力图, 能够动态强化教师-学生特征中的关键区域 (如目标主体), 抑制背景噪声. PCSA 通过跨通道注意力矩阵建模通道间依赖关系, 识别判别性强的语义通道 (如类别敏感特征), 抑制冗余噪声通道. 该融合了空间语义信息与通道间依赖关系. 通过空间-通道协同注意力机制, 该特征能够有效增强关键区域响应, 同时抑制背景噪声等无关信息. 我们使用了 MSE 损失距离来衡量F_t'和F_s'. 总蒸馏损失如下:

$$F_s' = SCSA(F_s),\ F_t' = SCSA(F_t) \tag{18}$$

$$L_{attn} = \frac{1}{N}\sum_{i=1}^{N}(F_s' - F_t')^2 \tag{19}$$

$$L_{SCAR\text{-}KD} = L_{fea} + \alpha L_{attn} \tag{20}$$

其中, α是权重系数为超参数. 对于超参数的设置, 我们先是在{0.1, 1, 10, 100}这组常用的指数级数值中初筛出合适的范围, 然后在其周围微调尝试不同的值来选出最佳参数.

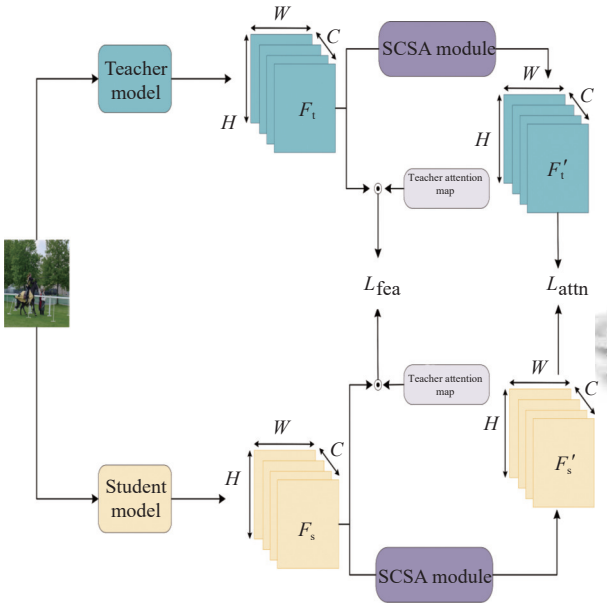


图 3 双重路径精炼特征蒸馏示意图

3 实验

3.1 实验数据集

我们在 VOC 和 VisDrone 公共数据集上验证了所提出的空间-通道协同注意力的精炼特征蒸馏 (SCAR-KD) 方法的有效性, 并将实验结果与其他 YOLO 系列框架进行了对比.

VOC 数据集由牛津大学的 Everingham 等人[24]的机器视觉团队于 2007 年发布. 首次发布的常用目标检测数据集, 主要用于小目标检测, 广泛应用于学术研究和工业实践. 该数据集包含 20 种不同的目标类别, 如车辆、动物、家居用品等, 数据标注精细且质量较高. 在实验过程中, 我们使用 VOC2007+VOC2012 训练集进行模型训练, 并在 VOC2007 测试集上评估模型性能, 以确保结果的公平性和可复现性. 整个数据集共包含 16 551 张训练图像和 4 962 张测试图像, 并提供了详细的边界框标注信息, 以支持目标检测任务的精确评估.

VisDrone 数据集则专为无人机场景下的目标检测任务设计, 特别适用于城市交通监控、安防巡逻和无人机导航等应用场景. 该数据集主要关注车辆 (如汽车、卡车、公交车等) 和行人目标, 涵盖了不同视角、高度、密度和尺度的目标分布, 具有极高的挑战性. VisDrone 数据集共包含 10 209 张图像, 其中包括 6 471 张训练图像、3 190 张测试图像和 548 张验证图像.

实验结果表明, 采用 SCAR-KD 蒸馏的模型在 VOC 和 VisDrone 数据集上均实现了显著的性能提升.

3.2 模型的训练及推理

在我们的实验中, 使用 YOLOv8n 作为学生模型, YOLOv8s 作为教师模型. 训练方法与 YOLOv8 保持一致, 包括使用带动量的随机梯度下降 (SGD)、分步衰减学习率以及预热策略. 为了提升训练性能, 采用两种强大的数据增强技术——Mosaic 和 Mixup, 这些方法均参考了以往的研究. 学习率设定为 0.01, 动量为 0.937, 权重衰减系数为 0.000 5. 完整的超参数列表可在 YOLOv8 代码中查阅. 所有实验结果均在两张 NVIDIA 3080Ti GPU 上训练得到. 推理时, 使用与 YOLOv8 相同的设置, 其中 IoU 阈值设定为 0.7, 图像输入尺寸为 640, 其他参数不再单独列出.

从图 4 中可以看出, 精炼后的特征图强调突出了图像中的关键区域, 如第 2 张图片中的自行车和第 3 张图片中的飞机群. 同时抑制了背景噪声如第 1 张图像中天上的电线和后面的树木. SCSA 先用不同大小的深度可分离卷积并行提取多尺度特征既有细粒度纹理, 也有大范围上下文. 再通过归一化和注意力融合, 只保留在任一尺度上对目标有用的信息, 抑制所有尺度都无意义的噪声, 得到一张既丰富又干净的空间注意力图. 学生只需对这张注意力图进行学习, 就能获取多尺度语义且不受背景干扰, 从而更快、更精确地贴合老

师的特征分布. 此外, SCSA 在此基础上再做通道自注意力, 使教师和学生在通道级别上的响应分布进一步对齐, 进一步缓解语义偏差.

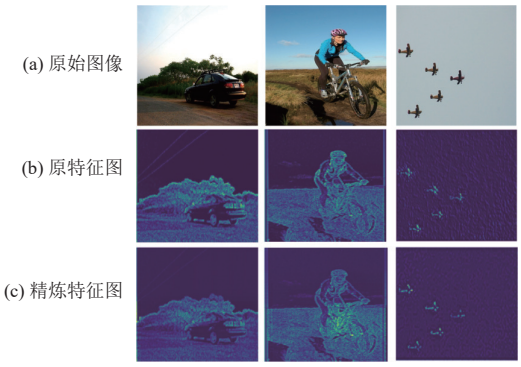
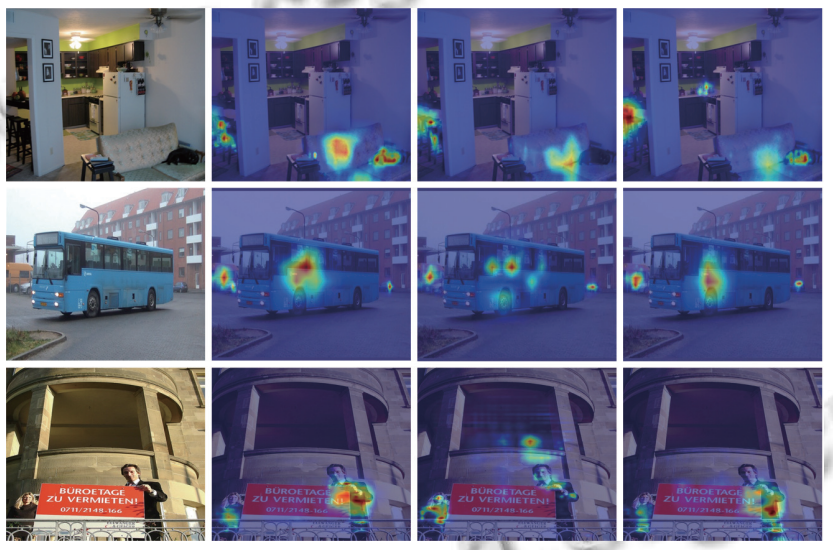


图 4 通过 SCSA 模块后的精炼特征图可视化

从图 5 可以明显看出, 采用本文方法的 YOLOv8n 模型在检测小目标及目标置信度方面有显著提升. 例如, 在第 1 张图片中, 右下角的小猫和桌子上的盆栽得到了更好的检测; 在第 2 张和第 3 张图片中, 应用我们方法的车辆和人体的热力值更加集中. 从图 6 可以看到, 第 1 张图片中黑暗背景下的水杯, 第 2、3 张图片海中的小船得到了更好的检测, 证明了采用我们方法的 YOLOv8n 模型在复杂背景或小目标的困难样本检测方面表现更优.

图 7 对比了使用 SCAR-KD 方法前后的训练效果曲线. SCAR-KD 蒸馏策略使模型在前期显著加快收敛并最终在 mAP@0.5 和 mAP@0.5:0.95 上提升约 1.2 个百分点, 同时训练曲线更平滑、对复杂场景更具鲁棒性.



(a) 输入图像 (b) YOLOv8s (教师) (c) YOLOv8n (学生) (d) YOLOv8n-SCAR-KD

图 5 教师模型、学生模型以及应用 SCAR-KD 方法后的学生模型热力图可视化对比

(a) YOLOv8n检测结果

(b) YOLOv8n-SCAR-KD检测结果

图 6 学生模型和使用了 SCAR-KD 方法蒸馏的学生模型检测结果对比

3.3 对比实验

在 VOC 数据集上, 本文针对不同模型的参数量 (Params)、FLOPs 以及 mAP 进行了比较, 结果如表 1 所示. 在表 1 中, mAP@0.5 代表 IoU 阈值为 0.5 时的累积均值平均精度 (mean average precision, mAP), 而输入尺寸表示输入图像的尺寸. 本文的主要关注点是 mAP@0.5:0.95 指标. 该指标综合考虑了 IoU 阈值从 0.5 到 0.95 范围内的检测性能. 需要说明的是, 由于早期方法 (如 SSD、CenterNet) 原生设计输入尺寸较小 (如 300、512), 本文未统一调整输入大小以保留其原始性能. 相比之下, YOLO 系列方法 (主要包括 YOLOv5、YOLOv6、YOLOv7、YOLOv8 及本文提出的 YOLOv8n-SCAR-KD)

均采用统一的 640×640 输入尺寸进行训练与测试, 确保了系列内部对比的公平性. 总体而言, 表 1 中的跨分辨率性能数据仅供参考, 本文主要关注 YOLO 系列在统一设置下的性能对比提升.

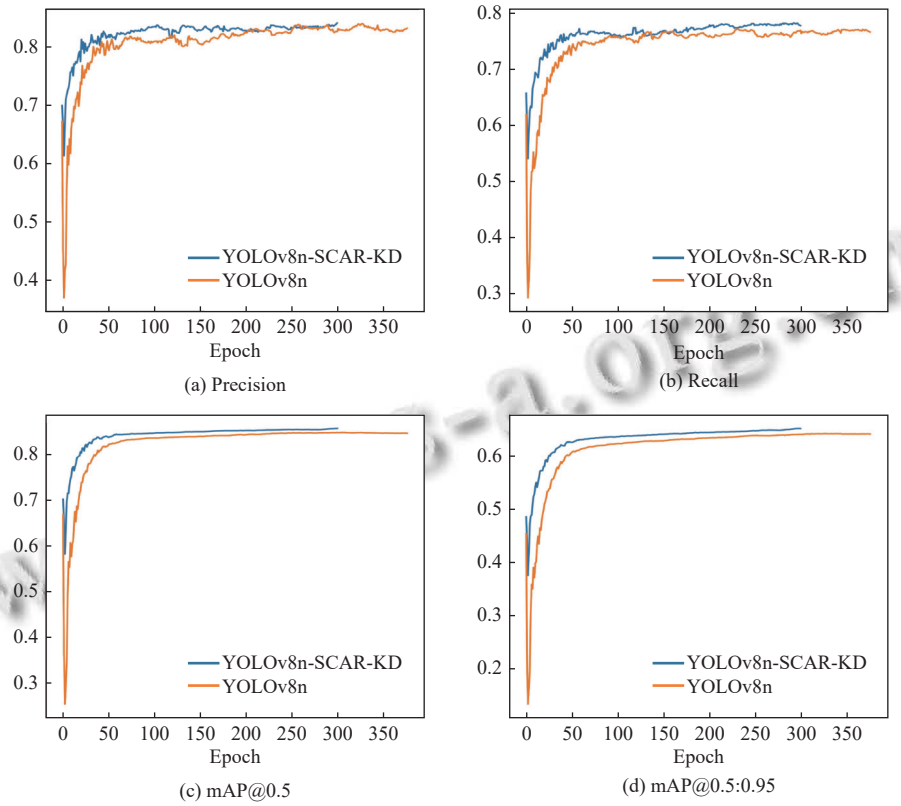


图 7 学生模型和 YOLOv8n-SCAR-KD 方法训练曲线图

表 1 在 VOC 数据集上不同模型性能对比

模型	输入尺寸	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Params (M)	FLOPs (G)
SSD[25]	300	74.3	—	26.3	62.5
CenterNet[26]	512	75.4	—	14.4	24.8
YOLOXT[27]	416	79.9	57.8	5.1	**6.5**
YOLOv5s	640	82.8	59.2	7.2	7.2
YOLOv6n[28]	640	84.8	63.1	4.3	11.1
YOLOv7t[29]	640	84.9	64.3	6.2	13.7
YOLOv8n[1]	640	84.1	64.1	3.0	8.1
YOLOv8n-SCAR-KD (ours)	640	**85.4**	**65.3**	**3.0**	8.1

在本文的实验中, 选择 YOLOv8n 作为学生模型, YOLOv8s 作为教师模型. YOLOv8n 在 VOC 数据集上的测试性能达到了 64.1%. 在引入 SCAR-KD 蒸馏方法后, 我们进一步将性能从 64.1% 提升至 65.3%. 在相同规模的网络中, 进行了不同蒸馏方法的对比, 我们的方法优于其他主流的蒸馏方法.

表 2 对比了不同知识蒸馏方法在 YOLOv8n 上的性能结果. 表 3 展示了使用 SCAR-KD 方法的 YOLOv8n 网络在 VisDrone 数据集上的测试结果对比.

表 2 在 VOC 数据集上同一模型下使用不同蒸馏方法对比 (%)

方法	蒸馏类型	mAP@0.5	mAP@0.5:0.95
YOLOv8n	—	84.1	64.1
CWD	feature	84.6	64.6
MGD[8]	feature	84.8	64.8
LD[30]	logit	84.7	64.7
BCKD[11]	logit	85.2	65.0
LWFI[5]	feature	85.0	64.8
YOLOv8n-SCAR-KD	feature	**85.4**	**65.3**

表 3 在 VisDrone 数据集上测试结果 (%)

方法	mAP@0.5	mAP@0.5:0.95
YOLOv8n	35.6	20.7
YOLOv8n-SCAR-KD	**36.8**	**21.6**

表 4、表 5 和表 6 分别展示了各超参数通过大范围初筛后得到的候选取值范围, 并在此基础上微调两个数据集超参数.

表 4 超参数粗筛阶段对比 (%)

α	VOC (mAP@0.5:0.95)	VisDrone (mAP@0.5:0.95)
0.1	65.12	21.33
1.0	65.19	**21.46**
10.0	**65.33**	21.42
100.0	64.77	20.92

表 5 在 VOC 数据集上超参数α对比 (%)

α	mAP@0.5	mAP@0.5:0.95
5.0	85.23	65.23
8.0	85.37	65.28
10.0	**85.42**	**65.33**
12.0	85.29	65.27

表 6 在 VisDrone 数据集上超参数α对比 (%)

α	mAP@0.5	mAP@0.5:0.95
2.0	35.43	21.49
2.5	35.45	21.50
3.0	35.55	21.52
4.0	**35.61**	**21.58**
5.0	35.52	21.49

3.4 消融实验

本文的消融实验主要集中在中间层蒸馏的层数选择以及各层损失的权重分配. 实验结果如表 7 所示.

表 7 在 VOC 数据集中依次加入蒸馏策略对比 (%)

方法	mAP@0.5	mAP@0.5:0.95
YOLOv8n	84.1	64.1
YOLOv8n+L_{fea}	84.9	64.8
YOLOv8n+L_{fea}+L_{attn} (SMSA)	85.1	64.8
YOLOv8n+L_{fea}+L_{attn} (PCSA)	85.3	65.1
YOLOv8n+L_{fea}+L_{attn} (SCSA)	**85.4**	**65.3**

如表 7 所示, 在 VOC 数据集中, 依次加入特征图动态加权蒸馏、空间多语义注意力模块 (SMSA) 精炼特征图、通道自注意力模块 (PCSA) 和空间-通道协同 (SCSA) 精炼特征图蒸馏后, 网络性能稳定提升.

在表 8 中, 当采用两个位置蒸馏时, 颈部位置的损失权重占比更大时, 性能提升更显著. 我们推测, 颈部位置的特征融合信息更加丰富, 因此需要学生网络更好地拟合. 在引入头部位置的特征蒸馏后, 我们同样采用之前工作中的使用 Softmax 自适应调整各位置的权重比例能够实现更稳定的性能提升.

表 8 VOC 数据集上特征蒸馏位置的权重选择 (%)

Backbone	Neck	Head	mAP@0.5:0.95
0.7	0.3	—	64.76
0.3	0.7	—	65.14
0.3	0.3	0.4	65.11
0.3	0.4	0.3	65.23
Softmax			**65.33**

4 结论

针对通过知识蒸馏提升轻量级网络性能这一目标, 本文提出一种 SCAR-KD 方法, 在原有工作的基础上, 我们设计了双路径的蒸馏框架, 在框架中应用了空间-通道注意力模块处理了原始特征图, 得到了精炼特征图中的关键区域并保留了丰富的语义信息. 最后, 所提出的 YOLOv8n-SCAR-KD 方法在 VOC 和 VisDrone 数据集上的实验验证显示出显著的性能提升, 证明了其在目标检测任务中的有效性.

参考文献

1 Varghese R, Sambath M. YOLOv8: A novel object detection algorithm with enhanced performance and robustness. Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). Chennai: IEEE, 2024. 1–6.

2 Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.

3 Sau BB, Balasubramanian VN. Deep model compression: Distilling knowledge from noisy teachers. arXiv:1610.09650, 2016.

4 Mirzadeh SI, Farajtabar M, Li A, *et al*. Improved knowledge distillation via teacher assistant. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020. 5191–5198.

5 Zhong LQ, Yan SY. Self knowledge distillation based on layer-wise weighted feature imitation for efficient object detection. Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul: IEEE, 2024. 9851–9855.

6 Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. Proceedings of the 5th International Conference on Learning Representations. Toulon: OpenReview.net, 2017.

7 Yang ZD, Li Z, Jiang XH, *et al*. Focal and global knowledge distillation for detectors. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 4633–4642.

8 Yang ZD, Li Z, Shao MQ, *et al*. Masked generative distillation. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 53–69.

9 Shu CY, Liu YF, Gao JF, *et al*. Channel-wise knowledge distillation for dense prediction. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision.

Montreal: IEEE, 2021. 5291–5300.
10 Park W, Kim D, Lu Y, *et al*. Relational knowledge distillation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3962–3971.
11 Yang LR, Zhou XP, Li XW, *et al*. Bridging cross-task protocol inconsistency for distillation in dense object detection. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 17129–17138.
12 Wang JB, Chen YM, Zheng ZH, *et al*. CrossKD: Cross-head knowledge distillation for object detection. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 16520–16530.
13 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.
14 Li X, Hu XL, Yang J. Spatial group-wise enhance: Improving semantic feature learning in convolutional networks. arXiv:1905.09646, 2019.
15 Park J, Woo S, Lee JY, *et al*. BAM: Bottleneck attention module. British Machine Vision Conference 2018. Newcastle: BMVC, 2018. 1–14.
16 Woo S, Park J, Lee JY, *et al*. CBAM: Convolutional block attention module. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 3–19.
17 Qin ZQ, Zhang PY, Wu F, *et al*. FcaNet: Frequency channel attention networks. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 763–772.
18 Ouyang DL, He S, Zhang GZ, *et al*. Efficient multi-scale attention module with cross-spatial learning. Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island: IEEE, 2023. 1–5.
19 Si YZ, Xu HY, Zhu XZ, *et al*. SCSA: Exploring the synergistic effects between spatial and channel attention. Neurocomputing, 2025, 634: 129866. [doi: 10.1016/j.neucom.2025.129866]
20 Yuan MY, Lang B, Quan FN. Student-friendly knowledge distillation. Knowledge-based Systems, 2024, 296: 111915. [doi: 10.1016/j.knosys.2024.111915]
21 Karine A, Napoléon T, Jridi M. Channel-spatial knowledge distillation for efficient semantic segmentation. Pattern Recognition Letters, 2024, 180: 48–54. [doi: 10.1016/j.patrec.2024.02.027]
22 Guo Z, Zhang PZ, Liang P. SAKD: Sparse attention knowledge distillation. Image and Vision Computing, 2024, 146: 105020. [doi: 10.1016/j.imavis.2024.105020]
23 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021.
24 Everingham M, van Gool L, Williams CKI, *et al*. The pascal visual object classes (voc) challenge. International Journal of Computer Vision, 2010, 88(2): 303–338.
25 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot multibox detector. Proceedings of the 14th European Conference on Computer Vision (ECCV 2016). Amsterdam: Springer, 2016. 21–37.
26 Zhou XY, Wang DQ, Krähenbühl P. Objects as points. arXiv:1904.07850, 2019.
27 Ge Z, Liu ST, Wang F, *et al*. YOLOX: Exceeding YOLO series in 2021. arXiv:2107.08430, 2021.
28 Li CY, Li LL, Jiang HL, *et al*. YOLOv6: A single-stage object detection framework for industrial applications. arXiv:2209.02976, 2022.
29 Wang CY, Bochkovskiy A, Liao HYM. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 7464–7475.
30 Zheng ZH, Ye RG, Wang P, *et al*. Localization distillation for dense object detection. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 9397–9406.

(校对责编: 张重毅)