计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009807] [CSTR: 32024.14.csa.009807]

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

基于键值注意力机制的目标检测算法性能优化①

张征鑫, 张笃振

(江苏师范大学 计算机科学与技术学院, 徐州 221116)
通信作者: 张笃振, E-mail: zhduzhen@jsnu.edu.cn

摘　要: 随着注意力机制在目标检测中的广泛应用, 进一步提升特征提取能力成为研究的重点. 提出了一种新的注意力机制, 旨在优化特征交互过程, 提升检测性能. 所提机制移除了传统自注意力中的查询操作, 采用深度可分离卷积高效提取局部与全局信息, 并通过键和值的加权融合实现特征聚合. 本文方法有效降低了计算复杂度, 增强了模型对重要特征的捕捉能力. 通过在 5 个不同类型的数据集上进行验证, 实验结果表明, 该注意力机制在处理小目标检测、遮挡处理以及复杂场景下的表现优异, 显著提高了检测精度与效率. 可视化分析进一步证实了其在特征提取中的有效性.

关键词: 目标检测; 深度学习; 注意力机制; 深度可分离卷积

引用格式: 张征鑫,张笃振.基于键值注意力机制的目标检测算法性能优化.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9807.html

Performance Optimization of Object Detection Algorithm Based on Key-value Attention Mechanism

ZHANG Zheng-Xin, ZHANG Du-Zhen

(School of Computer Science and Technology, Jiangsu Normal University, Xuzhou 221116, China)

Abstract: With the widespread application of the attention mechanism in object detection, further enhancing the feature extraction ability become the focus of research. A novel attention mechanism is proposed to optimize the feature interaction process and enhance the detection performance. The mechanism eliminates the query operation in traditional self-attention. It employs depth-separable convolution to efficiently extract both local and global information and realizes feature aggregation through the weighted fusion of keys and values. The method effectively reduces the computational complexity and enhances the model's ability to capture important features. Through validation on five different types of datasets, the experimental results demonstrate that the attention mechanism exhibits excellent performance in handling small target detection, occlusion processing, and complex scenes, significantly improving detection accuracy and efficiency. Visual analysis further verifies its effectiveness in feature extraction.

Key words: object detection; deep learning; attention mechanism; depthwise separable convolution

1　引言

在计算机视觉领域中, 检测并对图像内的目标进行定位和分类是一项关键性的工作. 目标的各种形态、尺寸、方向的差异, 加之光照条件的变化和遮挡等问题, 都令目标检测成为该领域面临的一大挑战. 深度学习技术的持续演进推动了目标检测的性能提升, 成为当前的主流趋势. 目标检测模型性能的提升主要聚焦于对数据的增强处理[1]、优化损失函数[2]以及网络架构的创新[3,4], 特别是注意力机制[5]的引入, 已在网络架构优化方面引起广泛关注. 注意力机制通过模仿

① 基金项目: 江苏师范大学研究生科研实践创新计划 (2024XKT2604)
收稿时间: 2024-09-13; 修改时间: 2024-10-30; 采用时间: 2024-11-01; csa 在线出版时间: 2025-02-18

人类的注意力分配过程, 使深度学习模型能够专注于输入数据的关键部分, 从而提升处理的精确度和效率. 它基于探索数据间内在的相关性, 进而强调图像的核心特征[6], 如空间注意力[7–11]、通道注意力[12–16]、混合注意力[17–23]、自注意力[24–27]和分支注意力[28–32]等, 目前已成为计算机视觉领域里应用最为广泛的技术之一.

尽管已有方法在许多方面取得了重要进展, 但仍存在几个问题.

(1) 空间信息处理不足: 现有注意力机制在强调特征通道的重要性时, 往往忽略了空间信息的充分利用. 这在需要精细空间定位的任务中可能影响模型的表现.

(2) 计算复杂度与轻量级网络兼容性: 许多高效的注意力机制因其复杂的设计和较高的计算需求, 在轻量级网络中的应用受限. 这在计算资源受限的环境下尤为突出.

(3) 跨层信息共享引起的信息过载: 尽管通过不同层级之间的注意力信息共享可以增强特征表达, 但这也可能导致信息过载, 使得模型难以专注于关键特征, 影响识别精度和泛化能力.

针对现有自注意力机制在计算复杂度和特征提取效率方面的不足, 提出了一种名为“键值注意力机制”的新型注意力机制. 该机制通过移除传统的查询 (Q) 操作, 利用深度可分离卷积提取局部和全局特征信息, 直接生成键 (K) 和值 (V) 以进行后续特征交互与加权. 具体来说, 键 (K) 用于表示图像中特征的模式, 即哪些特征是重要的, 而值 (V) 则表示与这些特征相关的具体内容信息. 通过卷积操作提取的 K 和 V, 代替了传统的 QKV 交互方式, 避免了查询步骤所带来的计算开销. 后续通过计算 K 和 V 之间的相似度, 实现对重要特征的权重分配和融合, 从而精确聚焦于关键区域.

实验结果表明, 键值注意力机制在多个公开数据集上显著提升了检测精度和效率, 展示了其在多种场景中的广泛适用性. 进一步的可解释性分析验证了该机制能够有效捕捉图像中的关键特征, 增强了模型的透明性, 为目标检测算法的优化提供了新的研究方向.

2 相关工作和背景

2.1 多种注意力机制分析

在深度学习研究中, 注意力机制被广泛认为是提升模型性能的关键技术, 主要分为硬注意力和软注意力两种形式. 硬注意力通过采样或选择的方式只关注输入的特定部分. 具体来说, 它会选择输入序列中的某些部分进行处理, 而忽略其他部分. 其优点是可以减少一些时间和计算成本, 但也可能丢失一些重要信息. 例如贝叶斯注意力[33]、期望最大化注意力[34]和高斯注意力[35]等. 软注意力是一种可微分的注意力机制, 它为输入的每个部分分配一个连续的权重值 (通常在 0–1 之间), 并使用这些权重来计算加权平均作为输出. 优点是它能够处理整个输入序列, 但计算量相对较大. SE、CBAM、ECA、坐标注意力、反向注意力、交叉注意力等是代表性的软注意力方法.

为了更直观地展现出各种注意力机制的特点, 表 1 总结了不同注意力机制的核心特点[12–14,17,24,36–47].

表 1 不同注意力机制特点比较

类型	注意力机制	特点
自注意力机制	SAM[24]	分析输入特征间的内在相关性, 通过自学习生成权重, 有效捕捉长距离依赖关系, 提升全局信息处理能力.
	Linear[37]	利用线性计算代替传统二次复杂度运算, 提高计算速度, 适合处理大规模数据.
	Sparse[38]	通过稀疏性技术选取重要的注意力权重, 降低计算和存储成本, 保留关键信息.
	Efficient[39]	通过优化内存使用和计算流程, 提高处理速度, 减少能耗.
	Dynamic[40]	动态调整网络结构或参数以适应输入数据的变化, 提升模型灵活性和适应性.
	Local[41]	在局部区域内聚焦计算注意力, 减少整体计算负担, 适用于局部特征重要的图像任务.
	X-linear[43]	利用线性变换扩展自注意力机制的可扩展性, 适用于图像和其他复杂数据结构.
	EA[47]	通过优化注意力模型的计算结构, 显著降低了模型的复杂度和计算需求, 使得处理大规模数据集更为高效.
通道注意力	SE[12]	自学习特征通道的重要性, 强调重要特征并抑制次要信息, 提升对重要视觉特征的敏感性.
	ECA[13]	不需要显式降维, 通过局部交叉通道相互作用, 提升模型效率和表现力.
	FCA[14]	通过频域分析强调通道特征, 增强模型对频域信息的敏感性.
	CLAM[42]	通过在不同层之间共享注意力信息来提高特征的表示能力, 促进深层特征和浅层特征之间的信息交流.
空间注意力	Axial[44]	通过在图像的每个轴向上独立应用注意力, 有效地捕捉全局上下文信息, 同时保持计算效率.
	CAT[45]	通过对单通道特征图中分隔的图像块间应用注意力, 实现全局信息的收集, 同时有效捕获局部信息.

表 1 不同注意力机制特点比较 (续)

类型	注意力机制	特点
混合注意力	CBAM[17]	综合考虑空间和通道注意力, 通过顺序模块化方式增强关键特征表达, 优化图像内容的全面理解.
	BAM[36]	结合通道和空间注意力, 以瓶颈层为目标, 通过简洁的设计增强网络中间层的表达能力.
	HaloNet[46]	通过结合局部窗口和稀疏全局注意力, 在保持计算效率的同时, 提升了信息聚合的能力.

2.2 深度可分离卷积和自注意力机制

深度可分离卷积 (DSConv)[48]通过将传统卷积操作分解为深度卷积和逐点卷积两个步骤, 显著降低了参数数量和计算成本. 具体来说, 深度卷积对每个输入通道独立应用卷积核, 有效提取空间特征; 而逐点卷积则通过 1×1 卷积核整合深度卷积的结果, 捕捉通道间的相关性. 其计算如式 (1):

$$O_{n,k,l}=\sum_{i,j,m}K_{m,n,i,j}\cdot F_{m,k+i-1,l+j-1} \tag{1}$$

其中, $O_{n,k,l}$表示输出特征图第n通道在位置k, l的值; $F_{m,k+i-1,l+j-1}$表示输入特征图第m通道在位置$k+i-1$, $l+j-1$的值; $K_{m,n,i,j}$表示卷积核在通道 m到n的卷积权重.

深度卷积对每个输入通道独立进行卷积, 如式 (2) 所示:

$$\hat{O}_{m,k,l}=\sum_{i,j}\hat{K}_{m,i,j}\cdot F_{m,k+i-1,l+j-1} \tag{2}$$

其中, $\hat{O}_{m,k,l}$表示深度卷积后输出的第m通道特征; $\hat{K}_{m,i,j}$表示 深度卷积核 (3×3) 在通道m上的权重.

逐点卷积将深度卷积的输出通过 1×1 卷积整合通道信息, 如式 (3) 所示:

$$O_{n,k,l}=\sum_{m}\hat{K}_{m,n}\cdot\hat{O}_{m,k,l} \tag{3}$$

其中, $\hat{K}_{m,n}$表示逐点卷积核从m通道到n通道的权重.

自注意力机制通过计算输入特征之间的相互关系来调整权重分布, 从而动态捕捉全局信息. 具体步骤如下.

首先, 将输入特征X通过 3 个不同的权重矩阵转换为Q、K和V, 如式 (4) 所示:

$$Q=XW^{Q},\ K=XW^{K},\ V=XW^{V} \tag{4}$$

接着, 通过计算查询与键的点积获取特征间的相似度, 并进行缩放与归一化, 如式 (5) 所示:

$$Attention(Q,K,V)=Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{5}$$

自注意力机制允许模型在处理每个特征时, 考虑与其他特征的全局关系, 从而增强对复杂场景的理解能力.

3 键值注意力机制

本研究提出了一种新的键值注意力机制, 旨在通过简化传统自注意力机制的计算流程, 提升模型在目标检测中的性能. 传统自注意力机制依赖查询 (Q)、键 (K) 和值 (V) 之间的交互来计算特征之间的全局相似度, 而这一机制在计算大规模数据时效率较低. 不同于传统方法, 本研究并非简单地用卷积替代查询步骤, 而是从根本上改变了特征提取与交互的方式.

在键值注意力机制中, 查询 (Q) 操作被完全移除, 取而代之的是利用深度可分离卷积直接生成键 (K) 和值 (V). 这里的卷积操作用于从输入数据中提取局部和全局信息, 而不是传统的 QKV 机制中的相似度计算. 这意味着模型不再需要通过 Q 与 K 的全局比较来获取特征的相似度, 而是通过对 K 和 V 之间的交互来直接进行特征聚合.

其中, 键 (K) 代表了特征模式, 表明输入数据中哪些特征是重要的; 而值 (V) 则表示与这些特征模式相关的具体信息. 深度可分离卷积能够有效提取输入特征的空间和通道信息, 生成的 K 和 V 在结构上更加精细, 适合局部特征的提取. 与传统的 QKV 交互不同, 键值注意力机制通过计算 K 和 V 之间的相似度, 实现了特征加权, 而不再依赖复杂的查询操作.

具体实现过程中, 输入特征首先通过深度可分离卷积进行处理, 提取局部的空间和通道信息. 深度可分离卷积不仅降低了计算复杂度, 还能更精细地捕捉局部依赖关系. 经过卷积处理后的特征通过线性变换生成键 (K) 和值 (V). 其中, K 表示输入特征中每个位置的模式, 而 V 表示与该模式关联的具体特征值. 尽管去除了 Q, 通过 K 和 V 的相似度计算, 我们仍能实现类似自注意力机制的特征加权. 如式 (6)、式 (7) 所示:

$$D_{n,k,l}=\sum_{m}\hat{K}'_{m,n}\cdot\left(\sum_{i,j}\hat{K}'_{i,j,m}\cdot F_{m,k+i-1,l+j-1}\right) \tag{6}$$

$$O=Softmax\left(\frac{W_K\cdot D^{\mathrm{T}}\cdot(W_V\cdot D)}{\sqrt{d_k}}\right)\cdot(W_V\cdot D) \tag{7}$$

其中, $D_{n,k,l}$表示卷积操作的输出特征图中的一个元素, 其索引为(n,k,l). n是输出特征图的通道索引, k和l 是空间位置索引. $\hat{K}'_{m,n}$是深度可分离卷积中使用的深度

卷积的卷积核, 表示第m个通道和第n个输出通道之间的深度卷积核权重. $\hat{K}'_{i,j,m}$是深度卷积核的元素, 表示在深度卷积中, 第m个输入通道的(i,j)位置的卷积核权重, $F_{m,k+i-1,l+j-1}$是输入特征图的第m个通道在位置$(k+i-1,l+j-1)$处的像素值. $W_K \cdot D^{\mathrm{T}} \cdot (W_V \cdot D)$计算的是键 ($K$) 和值 ($V$) 之间的相似度, W_K和W_V分别是线性变换矩阵, 将特征映射到不同的维度, 而 $\sqrt{d_k}$用于归一化防止数值过大, 通过点积操作, 计算了 K 和 V 之间的相似度. *Softmax* 函数进一步将相似度归一化为注意力权重, 这些权重用于对值 (V) 进行加权求和. 后面的乘法操作将这些注意力权重与 V 进行加权, 目的是让模型根据 K 的特征分布, 关注最相关的 V 信息, 从而完成最终的特征聚合和输出.

总的来说, 键值注意力机制通过去除查询并引入卷积操作, 简化了自注意力机制的计算过程. 卷积不仅有效提取了输入特征的局部依赖信息, 还确保模型能够高效处理局部特征, 并通过键和值的交互完成注意力计算. 这一设计提升了效率, 同时保留了对输入数据中关键特征的准确捕捉.

深度可分离卷积相较于传统查询操作的优势主要体现在以下几方面.

(1) 计算复杂度降低: 传统自注意力机制中, 每个查询向量需要与所有键计算得分, 计算复杂度为$\mathrm{O}(n^2)$. 而深度可分离卷积的复杂度为$\mathrm{O}(n)$, 计算复杂度大幅降低, 从而实现了计算上的高效性.

(2) 局部特征捕捉更加精细: 深度可分离卷积能够细致地捕捉和整合输入数据的局部与通道特征, 避免了传统查询机制可能造成的细节特征遗漏. 生成的键和值能够自然融合特征图中的局部与通道特征, 使得模型拥有更全面的特征视角.

(3) 减少过拟合的风险: 参数量减少意味着模型的自由度降低, 从而提高了模型的泛化能力, 使其能够更好地适应新的输入场景和变化.

图 1 是键值注意力机制模型结构图.

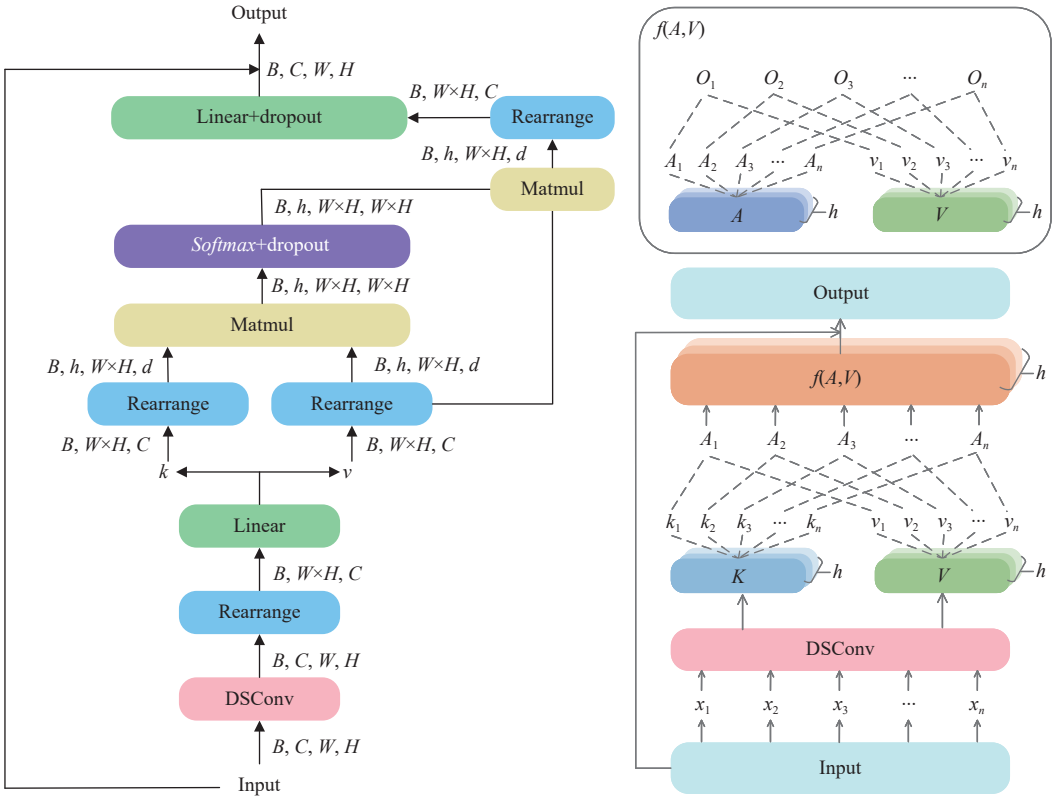

图 1 模型结构图

图 1 呈现了键值注意力机制的模型结构, 左侧部分展示了数据处理的流程: 输入数据首先通过深度可分离卷积 (DSConv) 处理, 在减少模型参数的同时, 更高效地捕捉局部特征, 增强模型对输入数据的理解能力. 卷积输出的特征经过线性变换, 形成键 (K) 和值 (V). 这些键和值分别表示了输入数据中的重要特征模

式和与这些模式相关的特征信息. 键和值被重新排列, 并通过矩阵乘法计算注意力权重, 随后再通过 *Softmax* 归一化. 这些权重反映了模型对不同特征的关注程度. 最终的加权值经过线性变换处理, 并通过残差连接与原始输入数据融合, 既保持了输入信息的完整性, 又强化了模型的特征表达能力, 从而生成最终的输出.

右侧部分则聚焦于模型的参数计算视角, 展现了键值注意力机制如何首先通过深度可分离卷积产生键和值, 然后通过它们的交互计算得到注意力权重. 这些权重与值相结合, 经残差处理后形成最终输出, 这一过程提升了模型对于重要特征的关注, 有助于在数据集上实现更准确的预测.

4 实验

为了全面评估键值注意力机制的性能, 以 YOLOv8 作为基准模型, 在第 8 层加入了键值注意力机制, 并在多个公开数据集上进行了实验. 选用的数据集特别关注小目标识别、目标遮挡处理以及日常场景下的目标检测, 以测试模型在处理这些环境下的有效性.

4.1 实验设置

本文实验基本环境设置如表 2 所示.

表 2 实验环境配置

Component	Specification
CPU	Intel Core i5-13500HX (32 GB)
GPU	NVIDIA GeForce RTX 4060 (8 GB)
System	Windows 11
Python	Version 3.8.17
PyTorch	Version 1.13.1
CUDA	Version 11.7

4.2 数据集概述

● PASCAL VOC[49]: 一个广泛用于目标检测的基准数据集. 结合了 PASCAL VOC 2007 和 PASCAL VOC 2012 的训练集与验证集, 以及 PASCAL VOC 2007 的测试集. 该数据集包含了 13 700 张图片作为训练集, 3 425 张图片作为验证集, 以及 4 952 张图片作为测试集. 涵盖 20 个类别的物体. 图 2 是 PASCAL VOC 数据集的类别柱状图.

● VisDrone2019[50]: 专注于无人机视角下的目标检测和跟踪, 包括大量小尺寸目标和密集场景. 其中包含 12 943 张训练集图片, 1 098 张验证集图片以及 3 221 张测试集图片. 涵盖 10 个类别的物体. 图 3 是 VisDrone-2019 数据集的类别柱状图.

● WiderPerson[51]: 专注于行人检测, 包括广泛的遮挡情形和复杂的人群场景. 该数据集包含了 16 002 张图片作为训练集, 2001 张图片作为验证集. 涵盖了 5 个类别的物体. 图 4 是 WiderPerson 数据集的类别柱状图.

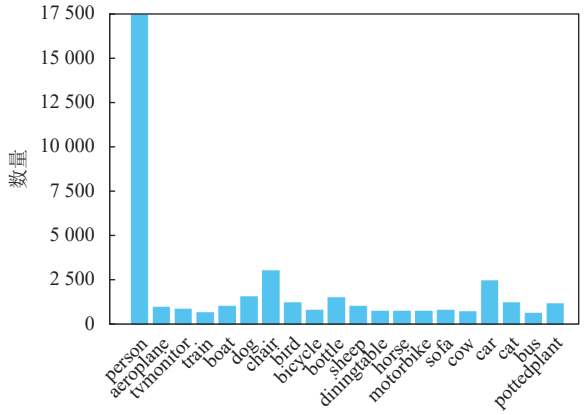

图 2 VOC 数据集类别柱状图

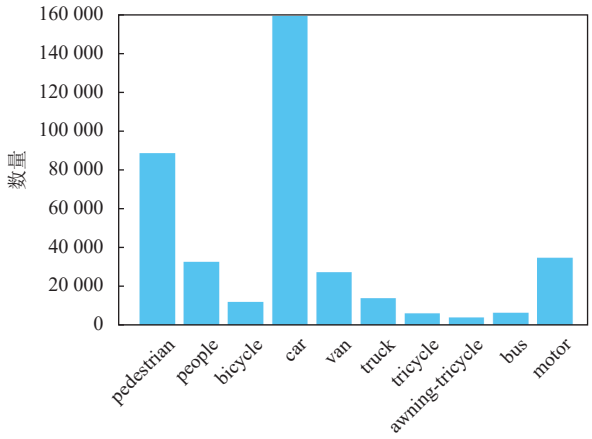

图 3 VisDrone2019 数据集类别柱状图

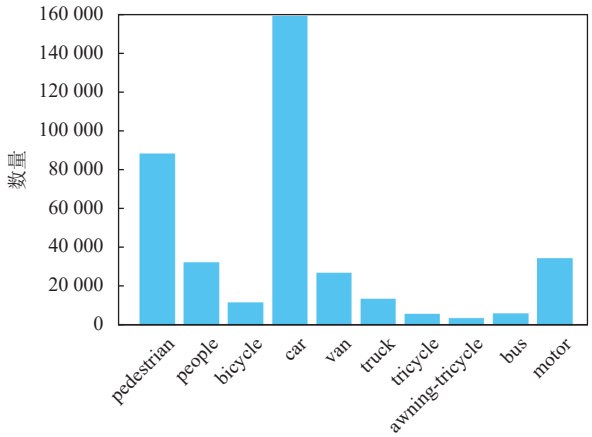

图 4 WiderPerson 数据集类别柱状图

●Vehicles: 车辆目标检测数据集, 其中包含 5 种常见车辆类型.

●TrafficSign: 交通标志数据集, 包括 30 种不同的交通信号标志.

4.3 评价指标

在对模型进行评估时, 通常考虑以下几个指标: 精确率 (*Precision*)、召回率 (*Recall*)、平均精度 (*AP*)、平均精度均值 (*mAP*) 和 *F*1 分数.

精确率反映了模型识别出的目标中, 正确识别的比例. 其计算公式为:

$$Precision = \frac{TP}{TP+FP} \tag{8}$$

其中, *TP* 表示真正例的数量, *FP* 表示假正例的数量.

召回率衡量的是模型识别出的目标占所有实际目标的比例. 其计算公式为:

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

其中, *TP* 表示真正例的数量, *FN* 表示假负例的数量.

平均精度是在不同召回率水平下, 精确率的平均值. 它是通过计算 *Precision-Recall* 曲线下的面积得到的. 以下是 *AP* 的计算公式:

$$AP = \int_0^1 Precision \, \mathrm{d}Recall \tag{10}$$

平均精度均值则是对所有类别的 *AP* 值进行平均, 提供了一个整体的性能指标. 其计算公式为:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{11}$$

其中, N是类别的总数, AP_i是第i个类别的平均精度.

*F*1 分数是精确率和召回率的调和平均数, 通常用于衡量模型的准确性, 特别是在类别不平衡的数据集上. 其计算公式为:

$$F1 = 2\times\frac{Precision\times Recall}{Precision+Recall} \tag{12}$$

*F*1 分数的值范围为 0–1, 其中 1 表示最佳可能的性能, 0 表示最差的性能. 一个高 *F*1 分数表明模型具有高精确率和高召回率, 平衡了两者之间的关系.

4.4 对比实验

4.4.1 实验 1

在第 1 个实验中, 重点探讨了加入键值注意力机制后的基准模型相较于其他目标检测模型的性能表现. 实验结果见表 3.

从表 3 中的数据中我们可以观察到, YOLOv8 作为当前流行的目标检测算法之一, 显示出了优异的性能. 因此, 本研究选择 YOLOv8 作为基准模型, 以此来探索和验证新的改进方法. 在实验中, 将键值注意力机制融入到 YOLOv8 的架构中, 结果表明, 此机制显著提高了模型的检测性能. 证明了此机制通过赋予模型以对关键特征的更高敏感性, 增强了对目标的表征能力, 使得在多个评价指标上有所提升.

表 3 经典目标检测算法在 VOC 数据集上性能对比 (%)

Model	*Precision*	*Recall*	*mAP*@50	*mAP*@50:95
Faster R-CNN[52]	69.7	57.8	62.3	41.2
SSD[53]	74.4	60.7	68.2	46.3
EfficientDet[54]	72.6	58.1	66.3	43.7
YOLOv5n	73.0	61.6	67.9	44.6
YOLOv8n	74.1	**65.6**	72.5	52.4
Ours	**76.1**	64.9	**73.1**	**53.0**

4.4.2 实验 2

在第 2 组实验中, 重点探讨了键值注意力机制相较于其他经典注意力机制的表现. 在 5 种数据集上的结果见表 4–表 8.

表 4 VOC 数据集 (%)

Model	*Precision*	*Recall*	*mAP*@50	*mAP*@75	*mAP*@50:95	*F*1
SE	75.8	64.6	72.7	56.9	52.4	69.1
CBAM	74.8	65.0	72.3	56.8	52.3	69.0
BAM	75.1	65.2	72.9	57.7	52.7	69.2
ECA	73.8	**65.9**	72.6	56.7	52.0	69.1
SAM	75.9	64.8	72.5	57.2	52.5	69.4
Ours	**76.1**	64.9	**73.1**	**58.1**	**53.0**	**70.1**

表 5 VisDrone2019 数据集 (%)

Model	*Precision*	*Recall*	*mAP*@50	*mAP*@75	*mAP*@50:95	*F*1
SE	36.2	25.1	23.5	12.4	12.7	28.3
CBAM	36.1	25.8	23.8	12.8	13.0	28.3
BAM	35.7	25.6	23.4	12.7	12.8	27.9
ECA	35.9	25.4	23.5	12.6	12.8	27.8
SAM	36.4	25.8	23.8	13.0	13.1	28.3
Ours	**36.7**	**26.1**	**23.9**	**13.2**	**13.3**	**28.5**

表 6 WiderPerson 数据集 (%)

Model	*Precision*	*Recall*	*mAP*@50	*mAP*@75	*mAP*@50:95	*F*1
SE	43.1	34.0	34.2	18.7	19.3	35.6
CBAM	43.5	34.2	33.8	**19.3**	19.1	35.5
BAM	43.1	**34.6**	34.0	18.9	19.3	33.9
ECA	43.2	34.1	34.5	18.8	19.3	35.5
SAM	43.6	34.2	34.0	18.6	19.1	35.1
Ours	**43.7**	**34.6**	**34.9**	18.6	**19.5**	**36.5**

从表 4–表 8 数据中可以看出, 相比于其他经典的注意力机制, 键值注意力机制在多个评价指标上都显示出了明显的提升. 这种提高证明了键值注意力机制

在目标检测中对细节的高度敏感性和精确度, 能够捕捉更丰富的信息, 以支持更复杂的决策过程. 该机制通过强化模型对关键局部特征的识别能力以及在全局上的信息融合能力, 显著提升了整体的检测性能. 在有效利用计算资源的同时, 通过细化对目标特征的处理, 不仅提高了模型的识别能力, 而且在实际应用中增强了系统的可靠性, 这对于提升技术的实用价值和广泛应用具有重要意义.

表 7 Vehicles 数据集 (%)

Model	*Precision*	*Recall*	*mAP*@50	*mAP*@75	*mAP*@50:95	*F*1
SE	83.1	62.1	66.2	58.0	52.6	65.3
CBAM	83.3	62.4	70.6	57.9	54.6	70.1
BAM	83.9	62.3	66.1	57.1	52.4	61.8
ECA	83.1	62.1	66.4	57.8	53.1	63.6
SAM	83.6	62.9	63.9	55.1	49.3	63.7
Ours	**84.0**	**63.3**	70.1	**58.4**	**55.7**	**70.3**

表 8 TrafficSign 数据集 (%)

Model	*Precision*	*Recall*	*mAP*@50	*mAP*@75	*mAP*@50:95	*F*1
SE	61.2	49.6	57.0	46.6	39.8	52.0
CBAM	61.4	49.9	59.2	48.8	**41.7**	53.3
BAM	61.0	49.2	54.9	45.3	38.5	48.7
ECA	61.1	49.5	56.3	45.5	40.0	52.5
SAM	61.2	49.3	**59.3**	48.6	**41.7**	53.4
Ours	**61.8**	**50.1**	58.8	**48.9**	**41.7**	**53.9**

4.4.3 实验 3

在第 3 组实验中, 重点探讨了键值注意力机制相较于其他改进的自注意力机制的表现. 在 5 种数据集上的结果见表 9–表 13.

表 9 VOC 数据集 (%)

Model	*Precision*	*Recall*	*mAP*@50	*mAP*@75	*mAP*@50:95	*F*1
Linear	75.1	64.2	72.6	57.3	52.5	69.1
Sparse	75.9	64.3	72.9	57.2	52.6	69.2
Efficient	73.7	65.9	72.5	57.0	52.4	69.1
Dynamic	74.8	64.2	72.5	57.0	52.3	69.1
Local	75.4	64.3	72.6	57.3	52.6	69.0
Ours	**76.1**	**64.9**	**73.1**	**58.1**	**53.0**	**70.1**

表 10 VisDrone2019 数据集 (%)

Model	*Precision*	*Recall*	*mAP*@50	*mAP*@75	*mAP*@50:95	*F*1
Linear	36.3	25.9	**23.9**	12.8	13.0	**28.6**
Sparse	36.4	25.8	23.7	12.7	12.9	28.1
Efficient	35.5	25.2	23.5	12.5	12.8	28.3
Dynamic	35.5	25.7	23.5	12.6	12.8	28.2
Local	36.4	25.6	23.4	12.5	12.8	28.2
Ours	**36.7**	**26.1**	**23.9**	**13.2**	**13.3**	28.5

从表 9–表 13 中的数据可以看出, 与其他改进的自注意力机制相对比, 键值注意力机制还是展现出了更加优越的性能. 传统的改进自注意力机制虽然能够有效地整合全局信息, 但通常在捕捉局部细节方面存在不足. 此外, 这些机制依赖于查询、键、值这 3 个参数的设置, 这不仅增加了模型的复杂性, 也有可能导致过拟合问题. 键值注意力机制通过使用深度可分离卷积专注于提取图像的局部特征, 然后将这些特征映射到键和值, 而非传统的自注意力机制中的查询、键、值三参数模型. 这种方法降低了模型的复杂性, 减少了参数的数量, 并专注于局部特征的识别. 使得键值注意力机制成为提升目标检测性能的一个有效工具.

表 11 WiderPerson 数据集 (%)

Model	*Precision*	*Recall*	*mAP*@50	*mAP*@75	*mAP*@50:95	*F*1
Linear	43.2	34.1	34.3	18.7	19.3	36.3
Sparse	43.3	34.1	34.2	18.9	19.4	35.1
Efficient	43.1	34.3	34.7	**19.4**	19.3	36.3
Dynamic	43.1	34.5	34.1	19.0	19.3	34.8
Local	43.5	34.4	34.4	19.1	19.4	35.3
Ours	**43.7**	**34.6**	**34.9**	18.6	**19.5**	**36.5**

表 12 Vehicles 数据集 (%)

Model	*Precision*	*Recall*	*mAP*@50	*mAP*@75	*mAP*@50:95	*F*1
Linear	83.5	62.2	68.6	**60.0**	54.1	69.3
Sparse	83.2	62.3	65.1	53.0	50.9	65.5
Efficient	83.1	63.1	67.5	56.3	53.6	64.9
Dynamic	83.6	62.5	63.8	53.0	49.4	65.5
Local	83.7	62.6	66.2	56.9	52.6	66.2
Ours	**84.0**	**63.3**	**70.1**	58.4	**55.7**	**70.3**

表 13 TrafficSign 数据集 (%)

Model	*Precision*	*Recall*	*mAP*@50	*mAP*@75	*mAP*@50:95	*F*1
Linear	61.4	49.9	58.8	48.3	42.0	55.2
Sparse	61.2	49.2	57.8	49.0	41.2	52.6
Efficient	**61.8**	49.3	59.2	49.0	41.6	53.5
Dynamic	61.3	49.8	56.1	46.6	39.9	53.2
Local	61.5	49.8	**59.9**	**50.7**	**43.1**	55.0
Ours	**61.8**	**50.1**	58.8	48.9	41.7	**53.9**

4.4.4 实验 4

下面的实验对多种注意力机制的性能进行了对比, 结果如表 14 所示.

表 14 不同注意力机制的性能指标对比

Model	Parameters	GFLOPs	FPS (f/s)
SE	3 017 740	8.1	686
CBAM	3 017 839	8.1	678
BAM	3 014 781	8.1	677
ECA	2 878 479	8.0	682
SAM	3 093 004	8.2	635
Linear	3 093 004	8.2	652
Sparse	3 093 004	8.2	383
Efficient	3 109 516	8.2	636
Dynamic	3 128 092	8.2	619
Local	3 109 516	8.2	367
Ours	3 076 620	8.2	655

从表 14 的数据分析可以看出, 键值注意力机制在参数数量、计算复杂度和检测速度上均表现优异. 尽管相比经典的通道空间注意力机制 (如 SE、CBAM) 稍增加了一些复杂性, 键值注意力机制更有效地融合了局部与全局特征, 优化了特征提取. 与其他改进的自注意力机制 (如 Local) 相比, 尽管后者在某些性能指标上略胜一筹, 但其较低的 FPS 反映出较高的计算成本. 总体而言, 键值注意力机制在确保较低的参数数量和较高的处理速度的同时, 提升了目标检测的整体性能, 使其在需要快速精确检测的应用场景中具有显著优势.

5 模型性能的可视化评估

为了深入探索键值注意力机制对模型性能的影响, 下面再采用曲线分析和热力图可视化的方法. 通过这些分析手段, 我们旨在展示键值注意力机制如何具体改进模型在处理复杂视觉任务时的表现, 尤其是在增强模型识别准确度和加速响应过程方面的能力.

5.1 曲线比较

曲线分析专注于评估键值注意力机制如何影响模型在关键性能指标上的表现. 通过观察这些指标随模型参数调整的变化曲线, 可以揭示键值注意力机制对模型性能的具体影响. 这一分析不仅有助于理解键值机制如何优化模型的整体表现, 还能突出其在不同操作条件下的适应性和效率. 在不同数据集上的结果如图 5–图 7 所示. 图中横坐标代表训练的轮次, 纵坐标代表检测精度, 分别显示的是 *mAP*@50 和 *mAP*@50:95 的值.

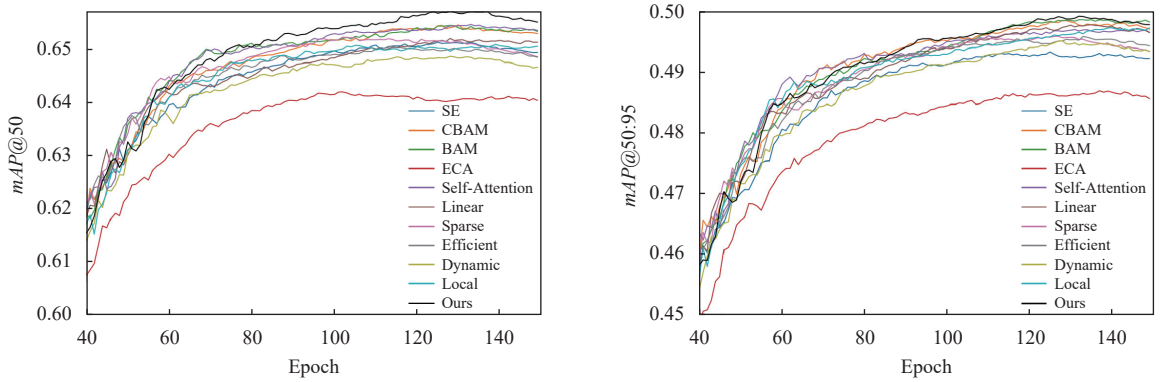

图 5 VOC 数据集模型性能比较图

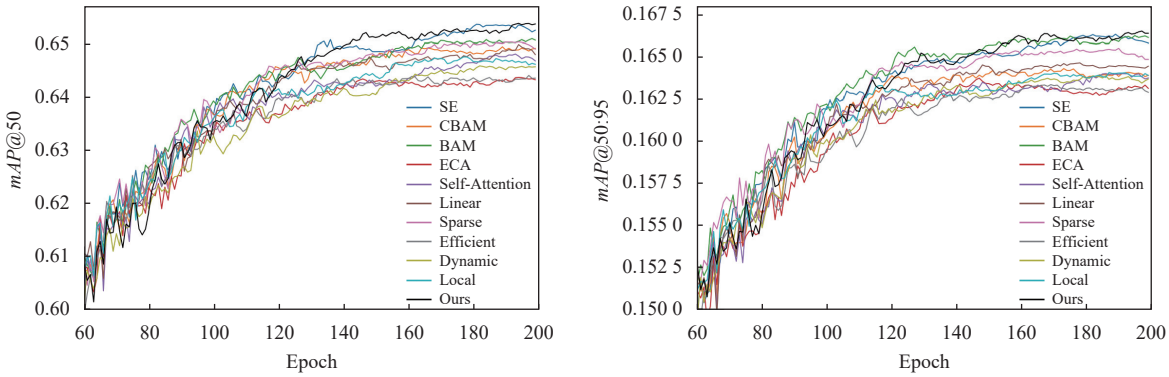

图 6 VisDrone2019 数据集模型性能比较图

通过深入分析键值注意力机制与其他常用注意力机制 (在 *mAP*@50 和 *mAP*@50:95 指标上) 的性能对比, 发现键值注意力机制在大部分训练过程中呈现出更快的学习速度和更高的准确率, 尤其是在训练进程的中后期, 展示了其在复杂图像目标检测任务上的卓越能力. 这一发现进一步证明了键值注意力机制在目标检测领域的实用性和优势. 然而, 值得注意的是, 在训练的初期阶段, 模型性能的提升不够显著, 指出了未来研究的方向, 即如何加强模型在早期训练阶段的学习效率. 总的来说, 键值注意力机制在应对多种挑战时展现了卓越

性能, 验证了其对未来研究和广泛应用前景的重要意义.

5.2 热力图对比

热力图分析则提供了一种直观方式, 展示模型在加入键值注意力机制后对图像各区域的关注程度. 通过比较不同注意力机制下模型的热力图 (图 8, 图 9), 可以直观地看到键值注意力机制如何使模型更加集中于图像中的关键信息, 从而提高识别的准确性. 热力图不仅展现了模型注意力的分布情况, 也反映了键值注意力机制在减少背景干扰和增强目标特征识别方面的有效性.

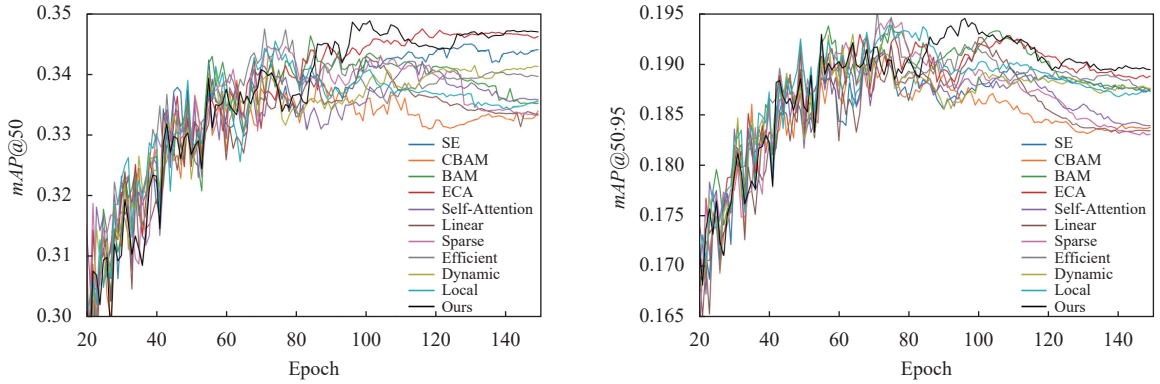

图 7 WiderPerson 数据集模型性能比较图

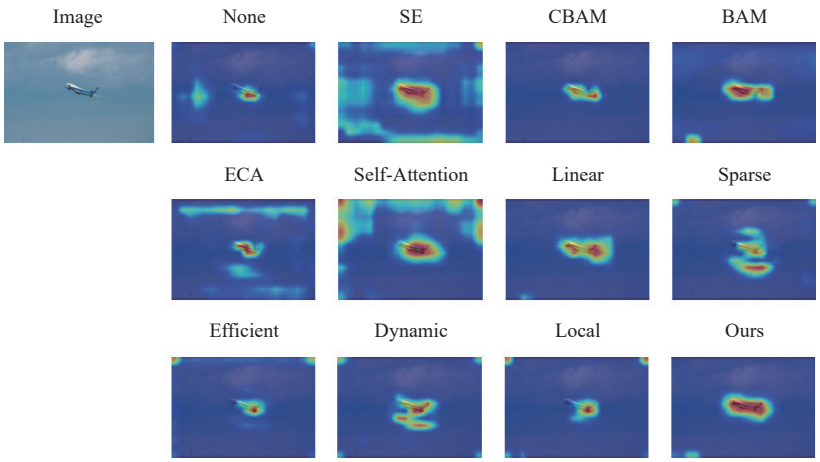

图 8 飞机图片不同注意力机制热力图比较

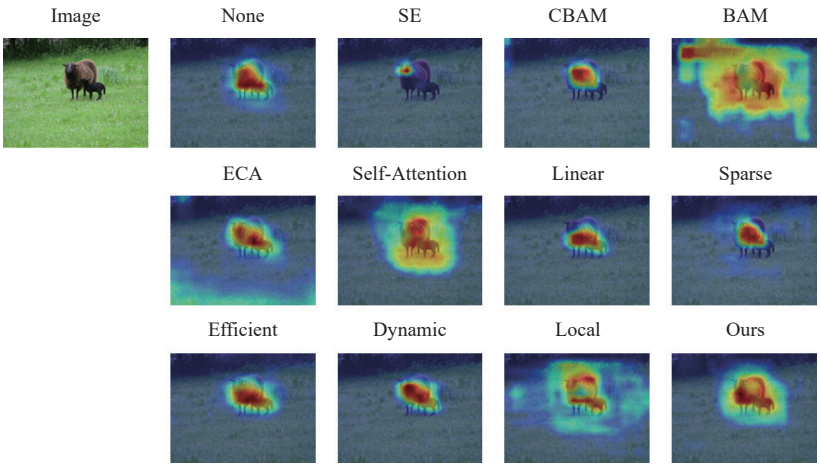

图 9 羊图片不同注意力机制热力图比较

图 8 和图 9 展示了在不同图像上应用多种注意力机制进行的热力图分析. 这些热力图是通过 Grad-CAM[55]生成的. 在模型的第 8 层加入了各种注意力机制, 并在此层对热力图进行了对比分析. 结果显示, 我们提出的键值注意力机制不仅是可行的, 而且在某些场合下, 它的表现甚至超过了其他的注意力方法. 此外, 从图中明显可以看到, 引入键值注意力机制之后, 模型的性能有了显著的提升.

6 总结

本文提出了一种新的键值注意力机制, 通过深度可分离卷积优化了特征提取和聚合过程. 传统自注意力机制依赖于复杂的查询操作来计算特征之间的全局相似度, 而本研究通过完全移除查询步骤, 利用卷积直接生成键和值, 简化了计算流程. 该机制有效地将局部和全局特征整合, 从而提高了模型对重要特征的捕捉能力. 实验结果表明, 该方法在多种目标检测任务中, 特别是在小目标检测、遮挡处理和复杂场景下, 均表现出色, 显著提升了检测的精度和效率. 综上所述, 提出的键值注意力机制为目标检测任务提供了高效的特征提取方案, 并在各种类型的数据集上展示了优异的性能.

参考文献

1 Zhang YZ, Wang WJ, Li ZM, *et al*. Development of a cross-scale weighted feature fusion network for hot-rolled steel surface defect detection. Engineering Applications of Artificial Intelligence, 2023, 117: 105628. [doi: 10.1016/j.engappai.2022.105628]

2 Zhang YF, Ren WQ, Zhang Z, *et al*. Focal and efficient IoU loss for accurate bounding box regression. Neurocomputing, 2022, 506: 146–157. [doi: 10.1016/j.neucom.2022.07.042]

3 Alferaidi A, Yadav K, Alharbi Y, *et al*. Distributed deep CNN-LSTM model for intrusion detection method in IoT-based vehicles. Mathematical Problems in Engineering, 2022, 2022: 3424819.

4 Guo ZQ, Xu LN, Si YJ, *et al*. RETRACTED: Novel computer-aided lung cancer detection based on convolutional neural network-based and feature-based classifiers using metaheuristics. International Journal of Imaging Systems and Technology, 2021, 31(4): 1954–1969. [doi: 10.1002/ima.22608]

5 Mnih V, Heess N, Graves A, *et al*. Recurrent models of visual attention. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2204–2212.

6 Hassanin M, Anwar S, Radwan I, *et al*. Visual attention methods in deep learning: An in-depth survey. Information Fusion, 2024, 108: 102417. [doi: 10.1016/j.inffus.2024.102417]

7 Schwartz I, Schwing AG, Hazan T. High-order attention models for visual question answering. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 3667–3677.

8 Zheng HL, Fu JL, Zha ZJ, *et al*. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5007–5016.

9 Wang XL, Girshick R, Gupta A, *et al*. Non-local neural networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7794–7803.

10 Hu H, Gu JY, Zhang Z, *et al*. Relation networks for object detection. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 3588–3597.

11 Zhang ZZ, Lan CL, Zeng WJ, *et al*. Relation-aware global attention for person re-identification. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 3183–3192.

12 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.

13 Wang QL, Wu BG, Zhu PF, *et al*. ECA-Net: Efficient channel attention for deep convolutional neural networks. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 11531–11539.

14 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.

15 Gao ZL, Xie JT, Wang QL, *et al*. Global second-order pooling convolutional networks. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3019–3028.

16 Diba A, Fayyaz M, Sharma V, *et al*. Spatio-temporal channel

correlation networks for action classification. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 299–315.

17 Woo S, Park J, Lee JY, *et al*. CBAM: Convolutional block attention module. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 3–19.

18 Hou QB, Zhou DQ, Feng JS. Coordinate attention for efficient mobile network design. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 13708–13717.

19 Fu J, Liu J, Tian HJ, *et al*. Dual attention network for scene segmentation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3141–3149.

20 Chen BH, Deng WH, Hu JN. Mixed high-order attention network for person re-identification. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 371–381.

21 Roy AG, Navab N, Wachinger C. Recalibrating fully convolutional networks with spatial and channel “squeeze and excitation” blocks. IEEE Transactions on Medical Imaging, 2019, 38(2): 540–549. [doi: 10.1109/TMI.2018.2867261]

22 Misra D, Nalamada T, Arasanipalai AU, *et al*. Rotate to attend: Convolutional triplet attention module. Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2021. 3138–3147.

23 Hou QB, Zhang L, Cheng MM, *et al*. Strip pooling: Rethinking spatial pooling for scene parsing. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 4002–4011.

24 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

25 Liu Z, Lin YT, Cao Y, *et al*. Pr Swin Transformer: Hierarchical vision Transformer using shifted windows. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 9992–10002.

26 Tan ZX, Wang MX, Xie J, *et al*. Deep semantic role labeling with self-attention. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018. 4929–4936.

27 Lin ZH, Feng MW, dos Santos CN, *et al*. A structured self-attentive sentence embedding. Proceedings of the 5th International Conference on Learning Representations. Toulon, 2017.

28 Yang B, Bender G, Le QV, *et al*. CondConv: Conditionally parameterized convolutions for efficient inference. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 117.

29 Zhang H, Wu CR, Zhang ZY, *et al*. ResNeSt: Split-attention networks. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 2735–2745.

30 Steiniger Y, Kraus D, Meisen T. Survey on deep learning based computer vision for sonar imagery. Engineering Applications of Artificial Intelligence, 2022, 114: 105157. [doi: 10.1016/j.engappai.2022.105157]

31 Li X, Wang WH, Hu XL, *et al*. Selective kernel networks. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 510–519.

32 Chen YP, Dai XY, Liu MC, *et al*. Dynamic convolution: Attention over convolution kernels. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 11027–11036.

33 Fan XJ, Zhang SJ, Chen B, *et al*. Bayesian attention modules. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1373.

34 Li X, Zhong ZS, Wu JL, *et al*. Expectation-maximization attention networks for semantic segmentation. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 9166–9175.

35 Zhang LW, Winn J, Tomioka R. Gaussian attention model and its application to knowledge base embedding and question answering. arXiv:1611.02266, 2016.

36 Park J, Woo S, Lee JY, *et al*. A simple and light-weight attention module for convolutional neural networks. International Journal of Computer Vision, 2020, 128(4): 783–798

37 Katharopoulos A, Vyas A, Pappas N, *et al*. Transformers are RNNs: Fast autoregressive Transformers with linear attention. Proceedings of the 37th International Conference on Machine Learning. PMLR, 2020. 5156–5165.

38 Child R, Gray S, Radford A, *et al*. Generating long sequences with sparse Transformers. arXiv:1904.10509, 2019.

39 Kitaev N, Kaiser Ł, Levskaya A. Reformer: The efficient Transformer. arXiv:2001.04451, 2020

40 Rao YM, Zhao WL, Liu BL, *et al*. DynamicViT: Efficient vision Transformers with dynamic token sparsification. Proceedings of the 35th International Conference on Neural Information Processing Systems. Curran Associates Inc., 2021. 1068.

41 Parmar N, Vaswani A, Uszkoreit J, *et al*. Image Transformer. Proceedings of the 35th International Conference on Machine Learning. Stockholm: PMLR, 2018. 4055–4064.

42 Li YY, Huang Q, Pei X, *et al*. Cross-layer attention network for small object detection in remote sensing imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 2148–2161. [doi: 10.1109/JSTARS.2020.3046482]

43 Pan YW, Yao T, Li YH, *et al*. X-linear attention networks for image captioning. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10968–10977.

44 Ho J, Kalchbrenner N, Weissenborn D, *et al*. Axial attention in multidimensional Transformers. arXiv:1912.12180, 2019.

45 Lin HZ, Cheng X, Wu XY, *et al*. CAT: Cross attention in vision Transformer. Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME). Taipei: IEEE, 2022. 1–6.

46 Vaswani A, Ramachandran P, Srinivas A, *et al*. Scaling local self-attention for parameter efficient visual backbones. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 12889–12899.

47 Shen ZR, Zhang MY, Zhao HY, *et al*. Efficient attention: Attention with linear complexities. Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2021. 3530–3538.

48 Chollet F. Xception: Deep learning with depthwise separable convolutions. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 1800–1807.

49 Everingham M, Eslami SMA, van Gool L, *et al*. The PASCAL visual object classes challenge: A retrospective. International Journal of Computer Vision, 2015, 111(1): 98–136. [doi: 10.1007/s11263-014-0733-5]

50 Du DW, Zhu PF, Wen LY, *et al*. VisDrone-DET2019: The vision meets drone object detection in image challenge results. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul: IEEE, 2019. 213–226.

51 Zhang SF, Xie YL, Wan J, *et al*. WiderPerson: A diverse dataset for dense pedestrian detection in the wild. IEEE Transactions on Multimedia, 2020, 22(2): 380–393. [doi: 10.1109/TMM.2019.2929005]

52 Ren S, He K, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(6): 1137–1149.

53 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot multibox detector. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 21–37.

54 Tan MX, Pang RM, Le QV. EfficientDet: Scalable and efficient object detection. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10778–10787.

55 Selvaraju RR, Cogswell M, Das A, *et al*. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 618–626.

(校对责编: 张重毅)