学校代码：10284
分 类 号：TP181
密　　级：公开
U D C：004.8
学　　号：MG21370021

南京大学

硕 士 学 位 论 文

论 文 题 目　基于卷积神经网络的视觉注意力机制研究

作 者 姓 名　卢俣金

专 业 名 称　计算机科学与技术

研 究 方 向　神经网络的设计与应用

导 师 姓 名　申富饶 教授

2024 年 5 月 28 日

答辩委员会主席 戴新宇 教授

评　　阅　　人 武港山 教授

徐明华 教授

论文答辩日期　2024 年 5 月 16 日

研究生签名:

导师签名:

Research on the Visual Attention Mechanism Based on Convolutional Neural Networks

by

Lu Yu-Jin

Supervised by

Professor Shen Fu-Rao

A dissertation submitted to

the graduate school of Nanjing University

in partial fulfilment of the requirements for the degree of

MASTER

in

Computer Science and Technology

School of Artificial Intelligence

Nanjing University

May 28, 2024

南京大学研究生毕业论文中文摘要首页用纸

毕业论文题目：基于卷积神经网络的视觉注意力机制研究

计算机科学与技术 专业 2021 级硕士生姓名：卢侯金

指导教师（姓名、职称）：申富饶 教授

摘　要

随着深度学习的发展，卷积神经网络（CNN）结构变得愈发复杂，提升了构建新型 CNN 的难度。因此，利用即插即用的组件来增强现有网络性能变得尤为重要。特别是在计算机视觉领域，引入注意力机制以聚焦于图像的关键区域，已被证实能有效提升网络性能。卷积核注意力机制根据输入动态调节卷积核权重，而特征图注意力机制则强调感兴趣的特征区域，两者都能提高 CNN 在特征提取上的效率，改善视觉任务的处理结果。本文针对以上两个维度，探讨了现有注意力的局限，并提出了两种互补的新型注意力机制，通过实验证明了它们的有效性。本文的主要研究内容与贡献如下：

1. 本文提出了一种基于局部特征的卷积核注意力机制 LADConv。首先，本文指出了传统方法在计算卷积核注意力系数时仅依赖特征通道信息的不足，并引入了一种新的策略：直接从卷积核覆盖的局部特征区域中提取注意力机制，随后利用自注意力深化局部特征之间的关系，最后引入位置注意力进一步调整注意力系数。在标准视觉任务数据集上的实验显示，LADConv 在不增加额外计算和参数的情况下，显著优化了 CNN 的性能，且通过消融实验验证了各模块的作用。

2. 本文提出了一种细粒度的分组特征注意力机制 GAM。考虑到卷积神经网络中深层特征图通道数的增加，传统的全局特征注意力机制可能会遗漏某些关键信息。为此，GAM 将特征分组处理以施加注意力机制，采用结构共享和避免降维的策略来降低计算和参数成本，并通过组间学习模块增强不同特征组间的互动，实现精细的特征调整。实验结果显示，GAM 在性能上优于传统特征注意力机制，且能与 LADConv 结合进一步提升网络性能。最后通过注意力可视化，直观展示了 GAM 的优势。

3. 本文将上述两种注意力以灵活的组合形式运用到实际的垃圾检测系统中，该系统集成了数据标注、模型训练、实时监控等完整的目标检测工作流程。系统中包含了多种注意力机制方案，确保了其在资源约束的设备上仍能实现最佳的检测效果。

关键词：注意力机制；计算机视觉；卷积核；特征图

南京大学研究生毕业论文英文摘要首页用纸

THESIS: Research on the Visual Attention Mechanism Based on Convolutional Neural Networks

SPECIALIZATION: Computer Science and Technology

POSTGRADUATE: Lu Yu-Jin

MENTOR: Professor Shen Fu-Rao

ABSTRACT

As deep learning progresses, the architecture of Convolutional Neural Networks (CNNs) has grown increasingly intricate. This complexity presents a challenge in developing new CNN models. In this context, integrating plug-and-play components into existing networks emerges as a crucial strategy for enhancing their performance. This is particularly significant in computer vision, where employing attention mechanisms to concentrate on critical image areas has shown to significantly boost network efficacy. These mechanisms, whether applied to convolution kernels or feature maps, dynamically adjust to inputs or highlight interest areas, respectively. They enhance CNNs' feature extraction capabilities and their proficiency in processing visual tasks. Our study delves into the existing attention mechanisms' limitations and introduces two innovative attention mechanisms, showcasing their effectiveness through empirical research. The key findings and contributions of our work are outlined as follows:

1. We introduce LADConv, an innovative convolution kernel attention mechanism that leverages local features. This approach addresses the shortcomings of conventional methods, which rely solely on feature channel information for determining convolution kernel attention coefficients. LADConv utilizes self-attention to deepen the interconnections among local features and employs positional attention to refine these coefficients further. Our experiments on standard visual task datasets indicate that LADConv significantly enhances CNN performance without necessitating additional computational resources or parameters. Ablation studies further affirm the utility of each component.

2. We propose the fine-grained Grouped Attention Mechanism (GAM), designed to navigate the complexities of deep feature maps in CNNs, where traditional global feature attention mechanisms may overlook crucial details. GAM applies attention in a grouped manner, employing strategies for sharing structures and avoiding dimensionality reduction to lessen computational and parameter burdens. It also boosts feature group interactions via an inter-group learning module, facilitating meticulous feature refinement. Experimental evaluations demonstrate GAM's superiority over existing feature attention mechanisms, particularly when used in tandem with LADConv. Additionally, attention visualization techniques vividly confirm GAM's effectiveness.

3. Our research applies these two attention types in a synergistic mix within a waste detection system, integrating comprehensive target detection workflows that include data annotation, model training, and real-time monitoring. This system accommodates various attention mechanism configurations, optimizing detection outcomes on devices with limited resources.

KEYWORDS: Convolutional Neural Network; Computer Vision; Convolution Kernel Attention; Feature Map Attention

目　　录

插图目录

表格目录

第一章　绪论

1.1　研究背景及意义

在当今时代背景下，随着高性能计算设备计算能力的提升，人工智能技术正在经历其历史上的第三次浪潮。针对过去在深度网络训练中遇到的梯度衰减问题，Geoffrey Hinton[1]在 2006 年采用逐层初始化和整体反馈的方法，成功训练出了深度信念网络，并首次提出深度学习 (Deep Learning) 这一概念。这一突破促使许多研究者投入到深度学习的研究中，推动了其在各研究领域的应用。目前，深度学习已成为人工智能领域内的一个重点技术，在自然语言处理 (Neural Language Process，NLP)[2-4]、计算机视觉 (Computer Vision，CV)[5-8]和时间序列分析 (Time Series Analysis，TSA)[9-11]等方向展现出强大的能力。深度学习的发展不仅在学术界获得了广泛关注，其影响力也逐渐扩散至人们日常生活的各个方面。例如，八年前，Google 公司开发的基于深度学习技术的围棋机器人 AlaphGo[12]接连战胜了包括李在石、柯洁在内的世界顶尖棋手，让人们认识到人工智能技术的强大。得益于计算芯片和存储单元的持续进步，研究人员能够利用更加丰富的计算和数据资源，进一步推动人工智能技术落地。2022 年 11 月，由 OpenAI 团队开发的人工智能聊天机器人 ChatGPT[13]再次引起广泛关注，这一创新产品不仅在近人类语言文本生成方面展现了展现了卓越能力，还能完成多种复杂的语言任务，并与用户进行高度自然的互动。目前，多个行业正在探索如何将人工智能技术应用于日常生活，旨在通过技术提升生活质量。

计算机视觉技术作为人工智能领域内一项较为成熟的技术分支，它的存在已经遍布人们的日常所及之处。这包括但不限于机场和车站的人脸识别系统，电商平台的图像搜索功能，以及新能源汽车的自动驾驶技术。在此背景下，注意力机制 (Attention Mechanism)[14-17]的引入被认为是计算机视觉领域的一次革命性创新。注意力这一概念来源于人类大脑处理信息的一种机制，其作用是帮助筛

选和聚焦于环境中的关键信息，同时忽略其他不相关的信息。当个体将注意力集中在某一特定刺激或任务上时，关注的对象会导致任务相关神经元的放电率增加并抑制不相关或干扰性信息的神经活动。计算机视觉中的注意力机制受到了人类视觉系统和大脑处理信息方式的启发。在实际的视觉任务中，如图像识别[14-15]、目标检测[18-19]或语义分割[20-21]等，注意力机制通过动态调整神经网络模型内部特征图的权重分配，实现了对输入数据的选择性关注。这意味着模型能够识别并优先处理对当前任务更为关键的视觉特征，同时减少对不相关或干扰性信息的关注。这一自适应权重分配过程不仅提高了特征提取的效率，还增强了模型对关键信息的表征能力，进而在多变的视觉环境中提升了模型的识别性能和鲁棒性。

注意力机制受到广泛关注的一个原因是其即插即用（Plug and Play）特性。这意味着它可以轻松地集成到现有的神经网络架构中，在不增加过多计算负担的前提下增强模型的性能，并且不需要对原始架构进行大幅修改。由于无需从头开始设计新模型，研究人员和开发者可以快速地通过添加注意力模块来探索和实验不同的改进策略，从而加速研究和应用开发过程。并且，这种灵活性意味着注意力机制可以被应用于各种不同的网络架构和任务中，如卷积神经网络 (Convolutional Neural Network，CNN)[22-23]、循环神经网络 (Recurrent Neural Network，RNN)[24]和 Transformer[17]等。无论是处理文本、图像还是多模态数据，注意力模块都能助力网络更精准地挖掘输入数据中的关键信息，并专注于那些对当前任务至关重要的部分。此外，注意力机制在提高模型可解释性方面也有显著作用。通过可视化注意力权重，研究人员和开发者可以直观了解模型决策过程，识别对预测结果有显著影响的输入信息。这增强了模型的透明度和可信度，特别是在医疗诊断、自动驾驶等要求高可靠性的应用场景中尤为重要。

综上所述，注意力机制的引入降低了模型优化的难度，为计算机视觉领域的研究人员和开发者提供了一种高效、灵活而且有效的工具。对于计算机视觉中注意力机制的原理和应用的理解与探索，已成为该领域的一个重要研究方向。

1.2 研究现状及挑战

1.2.1 研究现状

在人工智能的研究和应用领域中，注意力机制已证明了其重要性和广泛应用的潜力。本文专注于探讨计算机视觉中深度神经网络所采用的注意力机制。在过去的十年里，注意力机制已逐渐成为计算机视觉领域的关键技术，其发展历程可概括为四个主要阶段：

2014 年，Google 的 DeepMind 团队首次提出了一种循环注意力模型[24](Recurrent Attention Model，RAM)，标志着深度神经网络与注意力机制结合的开端。此模型利用强化学习 (Reinforcement learning，RL) 的方法，通过设置奖励机制来指导注意力焦点在特征图上的移动，从而逐步提取和处理关键特征信息，以提高图像分类任务的性能。Gregor[25]等人受此工作启发，使用变分自动编码器 (Variational AutoEncoder，VAE) 来模拟人类绘画过程，通过重复部分生成而不是一次正向传播的方式来生成单个图像，显著改善了生成模型的性能。Xu[26]等人在上述注意力的基础上加入了通过最大化近似变分下界或由强化学习训练得到的随机注意力机制，并将其融入到通用的图像描述模型结构中，取得了较好的性能。在这个阶段，循环神经网络是附加注意力机制的必要工具。

第二阶段的注意力机制的特点是使卷积神经网络能够更有效地面对空间变换，并集中资源处理对当前任务更为重要的图像区域。其中，最具有代表性的工作是空间变换网络 (Spatial Transformer Networks，STN)[27]。它克服了卷积神经网络在处理图像时仅限于平移不变性的局限，通过显式学习平移、缩放、旋转等空间变换，STN 使得网络能够主动关注任务相关的图像区域，提供了一种具有变换不变性的创新注意力机制。Dai[28]等人进一步扩展了这一概念，通过在有效感受野中引入可学习的偏移量，提出了可变形卷积 (Deformable Convolutional Networks)，从而实现了对图像重要区域的自适应选择，显著提升了目标检测和语义分割任务的性能。在此基础上，Zhu[29]借鉴知识蒸馏的思想，使用一个 R-CNN(Region-based Convolution Neural Networks, R-CNN) 作为教师网络指导训练过程，使网络能够更好地选择感受野的边界，从而减少不相关的区域对网络的干扰，增强了网络可变形采样的能力，能够关注更恰当的图像区域。

注意力机制发展的第三阶段主要是对特征图进行重塑，始于 SENet[14]。它

通过引入一个创新的压缩激励 (Squeeze-and-Excitation) 模块，赋予了网络自动调整特征图各通道权重的能力，旨在增强有效特征的同时抑制无关特征，以此提升网络的表征能力及计算效率。相比于 SENet 只关注特征图的通道重要性，CBAM 注意力机制[15]还引入对特征空间维度的考量。CBAM 将通道注意力机制和空间注意力机制解耦以提高计算效率，并通过引入全局最大池化和全局平均池化同时提取注意力系数，获得了更好的性能表现。ECA[16]使用一维卷积取代 SENet 中的降维操作来确定通道之间的交互，从而直接对权重向量和输入之间进行对应关系的建模，简化了注意力模型的复杂度。自此阶段起，设计注意力机制时，研究者们力求结构简单高效，旨在保持较低参数和计算成本的同时，实现显著的性能提升。

注意力机制的第四阶段标志着从对特征级别的关注转向卷积核层面。在传统的卷积神经网络中，一个普遍的做法是对所有输入使用相同的卷积核参数，这种方法忽视了不同样本间的差异性。为了提高网络对各种样本的适应性和表现力，常规做法是增加网络的深度或宽度，但这无疑会引入更多的计算负担。CondConv[30]通过为不同的输入样本生成特定的分支注意力系数，并将这些系数应用于多个静态卷积核上，通过加权组合产生动态调整的卷积核，实现了在较低额外计算成本的情况下增强模型的表现力和适应性。进一步地，ODConv[31]扩展了这一概念，将分支注意力机制应用于卷积核的所有维度，进而优化了卷积核对于关键特征区域的关注度，显著提高了各类卷积神经网络模型的准确性。

除上述提到的注意力机制之外，自注意力机制 (Self-Attention) 也在近年来受到广泛关注。Vaswani[17]首先提出了自注意力机制，该机制迅速在自然语言处理领域取得了突破性的成果[32-33]。Wang 等人[34]率先将自注意力引入计算机视觉领域，并提出一种能够捕获长距离依赖关系的非局部 (non-local) 网络，该方法在视频理解和目标检测任务上都取得了显著的效果。此后，一系列研究例如 EMANet[35]、CCNet[36]、GAMNet[37]和 Stand-Alone Network[38]等进一步优化了这一机制，提升了处理速度、结果质量以及模型的泛化能力。最近，基于纯自注意力机制的深度网络架构——视觉变换器 (Visual Transformers)[39-41]引发了广泛关注，这些研究展示了自注意力模型巨大的潜力，并为新型网络架构的设计提供了新的思路。

1.2.2 研究挑战

尽管注意力机制的研究领域已经诞生了诸多的模型架构，并且在众多任务中展示了其有效性，但这一领域仍旧活跃，不断经历重要突破和快速发展。当前的注意力机制方法仍存在一定的局限性，其生成注意力系数的基本原理及其作用机制的深层次理解仍然有尚未发掘的空白。本文将针对基于特征和基于卷积核的注意力机制展开讨论，主要存在以下几个技术难点：

1. 提取视觉关键区域特征的挑战：在视觉任务中，从庞大且复杂的视觉数据中准确且有效地辨认和抽取关键信息极为重要。在复杂场景下，目标对象可能由于尺寸、视角、遮挡或背景干扰等多种因素而难以识别，使得从关键区域提取视觉特征成为一大挑战。此外，如何有效整合来自不同层级和区域的特征信息，以提升注意力系数的表征能力，也面临诸多难题。

2. 注意力机制的泛化能力和可扩展性问题：注意力机制旨在能够广泛应用于各类常见的卷积神经网络架构，并且能够适应多样化的任务与环境变化，而不仅仅是在特定数据集上表现出色。这要求所设计的注意力模型不仅需要捕获到普遍的视觉特征，同时也需具备适应新颖、未曾见过的样本或场景的能力。并且，模型的设计要尽量做到简洁通用，才能够更好地即插即用，用于提升多种常见的卷积神经网络在视觉任务中的表现。

3. 性能与资源消耗之间的平衡难题：引入注意力机制时，需要在不显著增加计算负荷的同时，尽可能提升模型的性能。在一些实际应用场景，如移动设备和实时处理系统中，计算资源非常有限。这就要求注意力机制不仅要有效，还要足够简洁，以减少计算量和内存需求。然而，简化模型设计往往会牺牲模型的性能，寻找二者之间的最佳平衡点成为了一大挑战。

目前，通用的视觉注意力机制主要分为基于特征图和基于卷积核这两种形式。针对基于卷积核的注意力设计，现有研究通常倾向于将从特征图上得到的通道注意力系数通过简单的维度变换后直接应用到卷积核的不同维度上，这种做法可视作为直接使用 SENet 特征图注意力中的压缩激励模块来提取卷积核上的注意力系数，难以充分挖掘卷积运算的固有优势。而在设计基于特征的注意力机制时，研究者为了捕获整体特征的注意力，常通过全局池化来降维处理高维特征信息，这样的处理忽视了拥有大量通道数的特征图内部复杂的语义信息，

从而导致重要特征信息的丢失，存在一定局限性。为了解决上述问题，本文引入了针对卷积核和特征图的两种新型注意力机制，目的在于更好地提升卷积神经网络在视觉任务中的性能表现。

1.3 研究内容与贡献

本文集中讨论视觉注意力机制，深入分析了当前基于卷积核和基于特征的视觉注意力机制存在的问题，并提出了两种创新型注意力模型。在针对卷积核的注意力机制模型方面，本文通过捕获卷积核所覆盖的局部空间特征及其在特征图上的不同移动位置来构建注意力模型，在不增加过多计算负担的情况下提升了模型性能，实现了在卷积核注意力机制领域的突破。而在基于特征的注意力机制方面，本文采用了一种将通道维度精细化为批量维度的策略，通过这种细粒度的处理方式对特征施加注意力，避免了全局性的操作，并保留各批量维度之间的相互联系，实现了更精确的特征重构。此外，该特征注意力模型还可以与卷积核注意力模型相结合，以进一步提升卷积神经网络在视觉任务中的性能表现。本文还将这两种创新的注意力机制应用于实际场景中，验证了其有效性。研究内容概述如下：

1. 本文提出了一种局部感知卷积核注意力机制 (Local-Aware Dynamic Convolution Attention，LADConv)。其核心创新点在于对特征压缩模块的设计，该模块通过对卷积核在特征图上滑动过程中产生的局部特征进行捕获，作为构建注意力机制的基础。接着，利用自注意力机制来获得信息更加丰富的局部特征，并据此生成多维分布的注意力系数。随后，通过位置注意力机制来精准调整生成的注意力系数。特别地，当卷积核尺寸大小为 1 时，传统的卷积核注意力机制可视为 LADConv 的一个特例；随着卷积核尺寸的增加，LADConv 展现出更优越的局部特征捕获能力。本文进一步将 LADConv 融入到 ResNet[8]、MobileNetV2[42]和 ConvNeXt[43]等主流卷积神经网络架构中，并在多个流行视觉任务的标准数据集上验证了其性能。此外，通过丰富的消融实验，本文还强调了在卷积核注意力机制设计中捕获局部特征的重要性，进一步验证了该模型的有效性。

2. 本文还提出了一种细粒度的分组特征注意力机制 (Grouped Attention Module with Cross-Level Learning，GAM)。GAM 将特征图按通道维度分组重塑为多

个子特征，并行提取各子特征的通道注意力和空间注意力系数，再使用组间学习模块增强子特征注意力系数间的联系，随后将得到的注意力机制重新整合到各自的子特征上并恢复到原始的维度。GAM 的关键在于采用了更细粒度的策略对每个特征组内的通道和空间权重进行动态调整，同时通过跨组交互策略来提升特征组间的协同表达力，从而实现更细粒度的特征重塑。GAM 能够轻松集成到现有的通用卷积神经网络架构中，并在 CIFAR-100[44]、ImageNet1K[45]和 MS COCO[46]等广泛使用的视觉任务数据集上取得优异的性能表现。并且，GAM 能够与 LADConv 一并施加于卷积神经网络中，进一步提升网络在处理视觉任务时的性能。最后通过对 GAM 聚焦的图像关键区域进行可视化分析，本文清晰地展现了 GAM 注意力机制的实际作用和优势。

3. 本文将提出的基于卷积核与基于特征图的注意力机制集成至一套垃圾检测系统中，该系统针对通过边缘设备捕获的路面信息进行垃圾检测与分类。鉴于需部署在资源有限的边缘计算平台上，系统设计了多种注意力增强策略以适应不同硬件条件，保证算法优化在有限的资源下仍能有效运行。系统的目标检测模型基于 Yolov8s[47]架构，采用类 CSPDarknet[48]的网络结构作为 backbone，因此可以使用本文提出的注意力机制进行优化。全套系统实现了从数据标注、模型训练到最终部署的完整过程，允许用户根据实际硬件限制挑选最适宜的优化策略，以实现在资源受约束的环境下最大化检测模型的性能表现。

1.4 本文组织结构

本文一共包含六个章节，总体结构如图1-1所示，以下是每个章节的核心内容概述：

章节一为绪论，概述了计算机视觉领域中注意力机制设计的背景及其重要性，接着分析了视觉注意力机制发展的几个阶段及其各自的特性所在，指出了当前研究中面临的挑战，并概括了本文为应对这些挑战所进行的努力。

章节二为相关工作部分，该章节首先对视觉注意力机制进行了详细介绍并给出了其形式化定义。紧接着，该章节从基于卷积核和基于特征图两个维度切入，详细讨论了这两个领域内主要研究方法的理念，并对它们的优势和局限进行了评估。

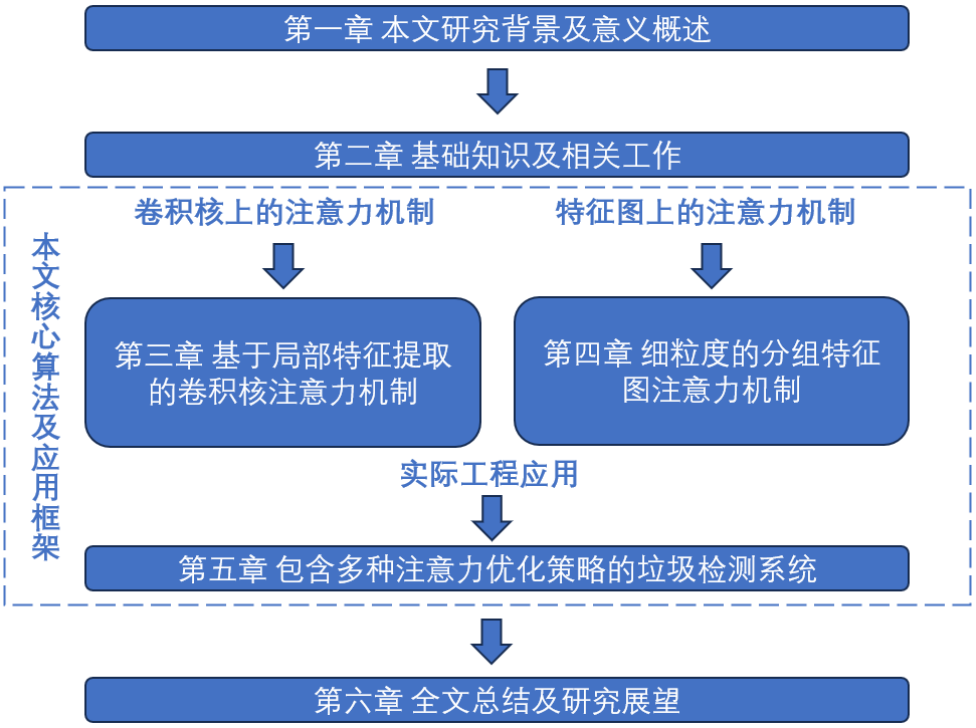


图 1-1 论文总体结构

章节三是对局部感知卷积核注意力机制的详细介绍。该章节首先阐述了 LADConv 提出的动机，随后详细解释了各个组成模块的具体设计。并通过一系列比较性实验展示了 LADConv 与其他技术相比的优越性。最后，通过设置消融实验论证了每个模块设计的合理性。

章节四是对细粒度的分组特征注意力机制的详细介绍。初始部分阐释了 GAM 的设计动机及实施要点，继而通过多项对照实验展示了 GAM 的性能表现。章节末尾，通过对注意力机制的视觉化呈现，提供了对 GAM 更为直观的理解。

章节五将本文中提出的两种视觉注意力机制应用于一个具体的垃圾检测系统中。该系统集成从数据标注到模型部署等全部流程，使用户可以根据实际的硬件条件需要，选择最合适的注意力机制充分挖掘目标检测模型在资源受限条件下的性能。

章节六为全文的总结和展望，概括了本文的研究内容及其贡献，并对所提方法的潜在局限和未来的改进方向进行了讨论。

第二章　相关工作

本章节主要介绍了实施视觉注意力机制所需的基础知识及关键技术，并概述了在此领域内一些具有里程碑意义的研究成果。首先介绍了注意力机制的定义以及施加注意力机制的关键模型——卷积神经网络，随后从两个关键方向展开讨论：一是基于卷积核的注意力机制，二是基于特征的注意力机制，逐一细述了几种在该领域内具有重要影响力的注意力设计思想。

2.1　注意力机制

人类的注意力机制是一种复杂的认知过程，使我们能够在众多环境刺激中，选择性地集中精力于某些特定的信息，同时忽略其他不相关的信息。通过一个简单的例子来体会视觉注意力机制的工作原理，如图2-1所示，人们在观察一幅图像时能够有效地分配有限的注意力资源：一开始看到这幅图时，观察者会立刻意识到这是一幅包含多位名人的黑白合影。然而，当观察者的视线集中到某个人物的面部具体细节上时 (以图中的爱因斯坦为例)，其他非关注区域则会被短暂地忽略。

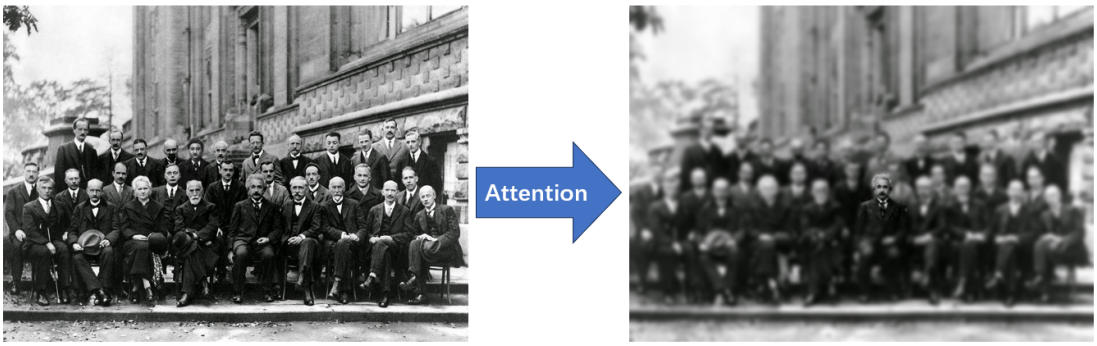


图 2-1　注意力机制的作用

在计算机视觉领域的研究中，注意力机制是一种使神经网络能够模仿人类视觉及认知系统的优化方式，它使神经网络在解析数据时能够更专注于关键信

息。通过引入注意力机制，网络有能力自主识别并优先处理输入数据中的显著特征，从而增强了网络模型的性能和泛化能力。将注意力机制进行形式化的定义，可表示为如下形式：

$$\text{Attention} = f(g(x), o), \tag{2-1}$$

其中，x 表示注意力的来源，在卷积神经网络中一般为卷积计算前后的特征图 (Feature Map)，$g(x)$ 表示根据来源产生注意力系数的过程，o 表示注意力的作用对象，通常为特征图或卷积核，$f(\cdot)$ 则表示注意力的施加过程。以 SENet 这种著名的注意力机制为例，$g(x)$ 和 $f(g(x), o)$ 可表示为公式2-2：

$$\begin{aligned} g(x) &= \text{Sigmoid}(\text{MLP}(\text{GAP}(x))), \\ f(g(x), o) &= g(x)o, \end{aligned} \tag{2-2}$$

这里的 MLP 表示多层感知机 (Multilayer Perceptron，MLP)，GAP 表示全局平均池化操作 (Global Average Pooling，GAP)，Sigmoid 是激活函数类型。

2.2 卷积神经网络

自 20 世纪 60 年代初，感知机 (Perceptron)[49]算法的提出标志着深度学习技术发展的起始阶段。尤其是卷积神经网络 (Convolutional Neural Networks，CNN)，自诞生之日起，就经历了迅速的发展。这一过程中，早期对哺乳动物视觉系统的神经生理学研究，揭示了大脑处理视觉信息的层次性机制，为后来的卷积神经网络模型分层设计提供了重要的理论基础。上世纪八十年代，神经网络研究领域取得关键进展，Kohonen 提出了自组织映射 (Self-Organizing Map, SOM)[50]这一种新的网络结构思想，虽然与卷积神经网络有所不同，但也反映了神经网络领域的多样化探索。而后，1998 年 LeCun[51]提出 LeNet-5 网络模型，将卷积层的概念成功应用于数字识别任务，这一成果标志着卷积神经网络在实践中的有效性得到了验证。21 世纪初，深度学习的发展加速，尤其是 2012 年的 AlexNet[5]网络在 ILSVRC 比赛中取得的突破性成就，极大地提高了图像识别技术的准确率，开启了深度学习在视觉任务应用的新时代。随后，多个先进的 CNN 架构诞生，如 Inception[7]网络引入了多尺度处理的思想，ResNet 则通过引入残差学习解决了深

层网络训练的难题。这些进步不仅限于视觉识别，卷积神经网络的应用领域也不断扩展，覆盖了视频分析、自然语言处理、医学图像分析等多个领域，展现了其在处理多维数据方面的强大能力。

卷积运算构成了卷积神经网络的核心。如图2-2所示，卷积核通过在特征图上滑动的方式，执行乘加操作，这一过程本质上是对局部特征的提取和分析。这种独特的操作是卷积神经网络具备平移不变性特征的基础，确保了模型对于输入数据中位置变化的鲁棒性。

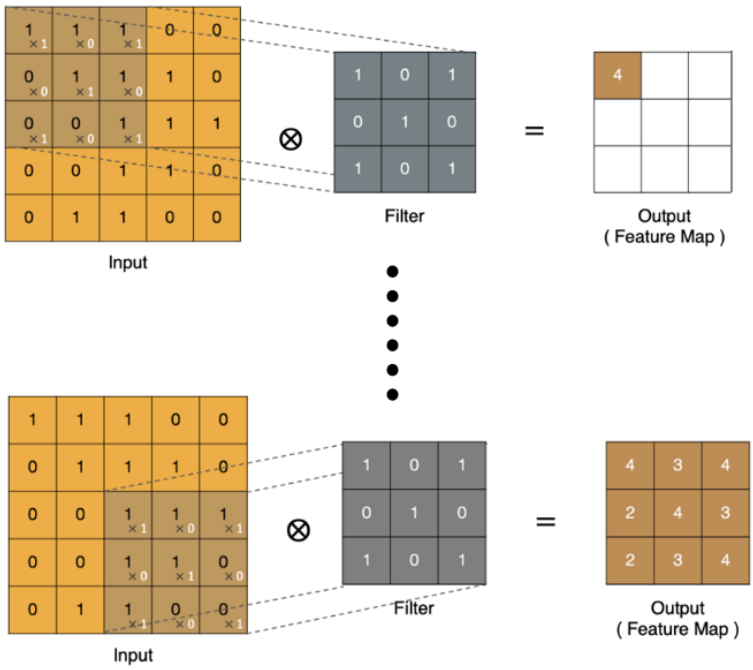


图 2-2 卷积计算示意图

接下来对卷积计算过程做出更形式化的定义。给定一输入特征图 $X \in \mathbb{R}^{H\times W\times C}$，$H$ 表示特征图的高度，W 表示特征图宽度，C 表示通道数。在一般情况下，卷积核的宽度和高度保持一致，这里都定义为 K，且卷积核的通道数等于特征图的通道数，所以卷积核可被定义为 $F \in \mathbb{R}^{K\times K\times C}$，假设一个卷积层有 N 个卷积核，卷积计算过程如下所示：

$$Y_{i,j,k} = \sum_{c=1}^{C}\sum_{u=0}^{K-1}\sum_{v=0}^{K-1} X_{i+u,j+v,c}\cdot K_{u,v,c}, \tag{2-3}$$

其中 $Y_{i,j,k}$ 表示输出特征图在第 k 个通道 (i,j) 位置上的值，$X_{i+u,j+v,c}$ 是输入特征图中与一次滑动操作所对应的相关元素的值，$K_{u,v,c}$ 是卷积核在第 c 个通

道 (u, v) 位置上的值。输出特征图的尺寸可以通过调整步长 S(Stride) 和边缘填充 P(Padding) 来控制。步长是卷积核在输入特征图上移动的距离。边缘填充是在输入特征图周围填充零的过程，用于控制输出特征图的空间尺寸。

2.3 基于卷积核的注意力机制

为卷积核施加注意力机制使静态卷积核参数动态化的过程已成为注意力机制研究中一项重要的创新。传统的卷积神经网络在训练完成后，其参数便固定不变，这意味着网络对所有输入样本均采用相同的参数进行特征提取，这种方法在处理多样化数据时显示出一定局限性。为了解决这一问题，研究者们提出了将注意力机制与卷积核相结合的方法。通过这种方法，网络能够根据不同输入样本产生特定的注意力权重，从而对卷积核的参数进行实时调整，以优化特征提取过程。这种参数动态化策略显著提升了模型在各种视觉任务上的性能。本节内容将详细介绍几种具有重要意义的卷积核注意力机制方法。

2.3.1 动态卷积

动态卷积(Conditionally Parameterized Convolutions for Efficient Inference，CondConv)[30]是首个将注意力机制运用于动态调整卷积核参数的工作，其思想简单且效果较为显著，根据输入特征生成权重注意力系数，然后将多个静态卷积核根据注意力系数加权得到一个参数能动态调整的卷积核。这种为卷积核施加注意力的思想给研究人员带来了很大的启发，后续出现一些工作都借鉴了 CondConv 的实现方式[52-53]。如图2-3所示，CondConv 背后的原理是使用多套静态卷积核，也称为多个“专家”，并通过一个路由函数生成注意力系数，这些系数决定了在当前输入样例下，哪些“专家”的贡献权重更为显著。这使得网络能够根据不同的输入自适应地调整所采用的卷积核权重，实质上构建了一个能够根据输入样本动态变化的卷积神经网络架构。可将 CondConv 表示为如下形式：

$$\text{Output}(x) = \sigma\left((\alpha_1 W_1 + \cdots \alpha_n W_n) * x\right), \tag{2-4}$$

其中，$\alpha_1, \cdots, \alpha_n$ 为生成的注意力系数，$W_1, \cdots, W_n$ 为静态卷积核的权重。

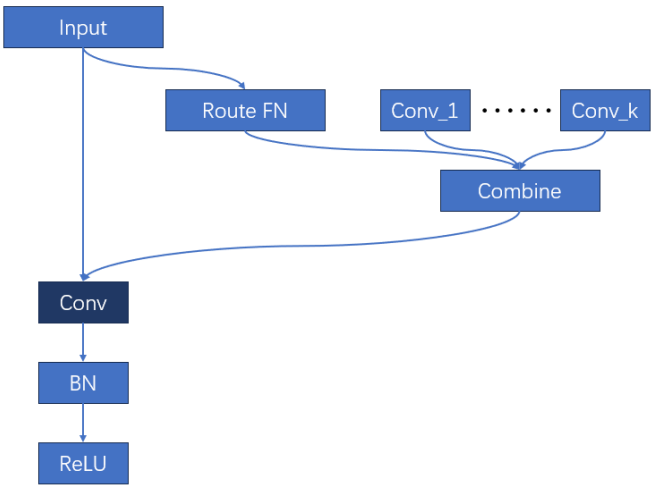


图 2-3 CondConv 结构示意图

在传统的卷积神经网络设计中，模型容量的扩展往往依赖于增加网络的宽度或者深度，这种方法虽然可以提升网络的表达能力，却也使得计算成本显著上升。与之形成鲜明对比的是，CondConv 采取了一种创新策略，通过计算每个输入样本特定的加权注意力系数，随后指导多个静态卷积核组合形成一个能够动态调整的卷积核。更为重要的是，得到的加权注意力系数可在同一层网络的不同卷积核上共享，避免了对每个卷积核进行重复计算的需求。因此，CondConv 仅仅实现了较少的推理时间增量，通过增加专家数量，能够有效提升网络的处理能力。CondConv 生成注意力系数的方式采用的路由函数可表示为如下形式：

$$r(x) = \text{Sigmoid}(\text{GAP}(x)R). \tag{2-5}$$

上式中，R 表示一个用于调整维度的多层感知机。不难发现，这种设计与 SENet 中生成特征图通道注意力的压缩激励模块高度相似。这意味着施加在卷积核上的注意力系数完全取决于输入特征图的通道属性，因此存在一定局限性。同时，引入大量静态卷积核会导致网络参数数量的显著提升，这不仅增加了训练网络时的计算负担，也可能在资源有限的边缘计算设备上造成部署难题。因此，虽然该方法在增强模型捕获特征的能力方面表现出一定的优势，但也需要权衡其对资源消耗的影响和在实际应用环境中的可行性。

2.3.2 矩阵分解动态卷积

将多个静态卷积核通过注意力加权，形成一个可动态调整的卷积核，虽然能够在一定程度上增强通用卷积神经网络在视觉任务中的表现，但这同样导致了网络参数量显著增加至原有的 k 倍 (这里的 k 取决于使用的静态卷积核数量)，此外，动态注意力机制与多静态卷积核的联合优化，也让网络训练过程变得更为复杂。矩阵分解动态卷积（Revisiting Dynamic Convolution via Matrix Decomposition，DCD）[54]重新评估了卷积核上注意力施加的方法，并从矩阵分解的角度深入探讨了其核心原理，发现动态化卷积核参数的根本在于将动态注意力机制映射到一个扩展的隐空间，随后针对特定通道组进行精确调节。为解决这一优化难题，DCD 提出了一种创新动态融合技术来取代传统的针对通道群体的动态注意力机制。这种动态融合方法不仅大幅度减小了隐空间的维度，还简化了模型训练的复杂度。因而，DCD 在减轻模型结构的同时，依旧维持了良好的性能，并且只需更少的参数量便可实现。

类似于 CondConv 这种为多个静态卷积核施加注意力聚合得到一个动态卷积核的方法可以形式化地表达为公式2-6：

$$W(x)=\sum_{k=1}^{K}\pi_k(x)W_k \quad \text{s.t.} \quad 0\leqslant\pi_k(x)\leqslant 1,\sum_{k=1}^{K}\pi_k(x)=1, \tag{2-6}$$

不难看出，最终聚合得到的卷积核是 k 个静态卷积核 W_k 基于注意力系数 $\pi_k(x)$ 的线性组合。对静态卷积核进行矩阵分解，可以将其重新定义为如下形式：

$$W_k=W_0+\Delta W_k, \quad k\in\{1,\cdots,K\}, \tag{2-7}$$

其中，$W_0=\frac{1}{K}\sum_{k=1}^{K}W_k$，可以视为多个静态卷积的平均。$\Delta W_k=W_k-W_0$ 是分解得到的残差矩阵。使用奇异值分解 (Singular Value Decomposition，SVD) 对残差矩阵 ΔW_k 进行进一步分解，可以得到 $\Delta W_k=U_kS_kV_k^{\top}$。因此，每个静态卷积核可以被重新表示为公式2-8：

$$W(x)=\sum_{k=1}^{K}\pi_k(x)W_0+\sum_{k=1}^{K}\pi_k(x)U_kS_kV_k^{\top}=W_0+U\Pi(x)SV^{\top}, \tag{2-8}$$

其中，$U = [U_1, \cdots, U_K], S = \text{diag}(S_1, \cdots, S_K), V = [V_1, \cdots, V_K]$。$\Pi(x)$ 是由注意力权重组成的矩阵 $\Pi(x) = \text{diag}(\pi_1(x)I, \cdots, \pi_K(x)I)$，$I$ 表示单位矩阵。假定卷积核的通道数为 C，经奇异值分解表示，为卷积核施加动态注意力的过程可以看作将输入投影到维度更高的空间中，即通过矩阵 U 将 C 个通道转化为 KC 个通道，再对高维空间的 KC 个通道施加动态注意力 $\Pi(x)$，最后通过矩阵 $SV^\top$ 还原至 C 个通道。因此，通过奇异值分解，可以发现这种施加注意力的方式实则是在高维空间中对通道施加注意力，在降维过程中可能导致一定的信息丢失，从而抑制相应通道的学习。公式2-8中的残差矩阵 $U\Pi(x)SV^\top$ 是一个 $C \times C$ 的矩阵，也可以将其看做是 KC 个秩为 1 的矩阵求和的结果，如公式2-9所示：

$$W(x) = W_0 + U\Pi(x)SV^\top = W_0 + \sum_{i=1}^{KC} \pi_{\lceil i/C \rceil}(x) u_i s_{i,i} v_i^\top, \tag{2-9}$$

在这种设计中，由于基本向量 u_i 和 v_i 在每一个秩为 1 的矩阵中都是独立使用，没有共享，这可能导致在多个静态卷积核之间存在冗余。此外，想要实现静态矩阵 U,V 与动态注意力机制 $\pi(x)$ 的同时优化存在一定难度，主要是因为较低的注意力权重有可能会对会限制 U,V 的学习过程。

为了有效减少高维空间中的通道数量，研究者们提出了一种创新的注意力施加方法 DCD，它采用动态通道融合技术以克服传统注意力机制施加方法的限制。这一策略通过一个全矩阵 $\Phi(x)$ 来实施。其中 $\Phi(x)$ 内的元素 $\phi_{i,j}(x)$ 均依赖于输入 x。此处，$\Phi(x)$ 构成了一个 $L \times L$ 的矩阵，旨在在潜在空间内对通道进行动态合并。该方法的核心理念在于大幅度降低潜在空间的维度，以构建更加紧凑的模型，即 $L \ll C$。动态通道合并的实施方式如式2-10所示：

$$W(x) = W_0 + P\Phi(x)Q^\top = W_0 + \sum_{i=1}^{L}\sum_{j=1}^{L} p_i \phi_{i,j}(x) q_j^\top. \tag{2-10}$$

在上述公式中，$Q \in \mathbb{R}^{C \times L}, P \in \mathbb{R}^{C \times L}$ 分别用于降低和还原各静态卷积核的维度，$\Phi(x) \in \mathbb{R}^{L \times L}$ 用于对潜在空间内的通道信息进行合并，并将其维度限制在 $L^2 < C$ 的范围内，以此实现参数量的显著减少。同时，DCD 策略也简化了联合优化的挑战，由于矩阵 P 和 Q 中的各列与大量注意力系数相连，P 和 Q 在训练中不会过度受到某些较小注意力值的影响，从而优化了模型的学习过程。

2.3.3 多维动态卷积

DCD 通过采用矩阵分解策略来优化注意力机制的作用方式，实现了网络参数数量和训练难度的有效降低，但其对网络性能带来的提升却稍显逊色。受到 CondConv 启发，Intel 的研究团队提出了多维动态卷积（Omni-Dimensional Dynamic Convolution，ODConv）[31]，这是一种在卷积核的所有维度上实施注意力机制的方法。在一定程度上，ODConv 可以视为 CondConv 的高级形式，它不仅继承了 CondConv 基于注意力机制整合多个静态卷积核的策略，还拓展到了卷积核空间、输入及输出通道等多维度的动态调整，从而定义为一种全维度的注意力施加策略。显而易见，由于其对卷积核动态化能力的全面提升，ODConv 即使在仅作用于单个卷积核的场景下 (不依赖于将多个静态卷积核聚合成动态卷积核的方法)，也能呈现出与 CondConv 相媲美甚至更优的性能表现。

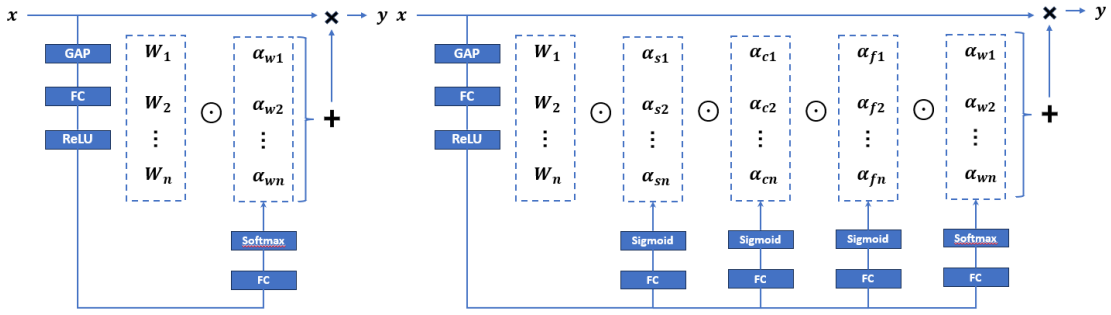


图 2-4 CondConv 和 ODConv 网络结构示意图

如图2-4所示，ODConv 保留了 CondConv 中路由函数和“专家”静态卷积核的设计，将基于特征的通道信息所产生的注意力系数通过全连接层 (Fully Connected Layer，FC) 进行对应的维度变换后，分别施加在各个静态卷积核的空间、输入通道与输出通道维度。延续公式2-4对 CondConv 的形式化定义，ODConv 可以被描述为如下形式：

$$y = \left(\alpha_{w1} \odot \alpha_{f1} \odot \alpha_{c1} \odot \alpha_{s1} \odot W_1 + \cdots + \alpha_{wn} \odot \alpha_{fn} \odot \alpha_{cn} \odot \alpha_{sn} \odot W_n\right) * x. \tag{2-11}$$

其中，α_{wi} 是用于聚合多个静态卷积核的注意力机制，$\alpha_{si} \in \mathbb{R}^{k\times k}, \alpha_{ci} \in \mathbb{R}^{c_{in}}, \alpha_{fi} \in \mathbb{R}^{c_{out}}$ 则是新加入的静态卷积核空间维度、输入通道与特征输出通道维度 (同一卷积层上有多个卷积核，它们的数量决定了特征的输出通道数)，这四种维度的注意力并行产生，前者的激活函数为 Softmax 函数，后三者的激活函数

统一采用 Sigmoid 函数。理论上，这四种不同类型的注意力机制相互补充，通过逐步将注意力沿着位置、通道、同一卷积层内卷积核数量以及专家静态核等多个维度应用于卷积操作，可以赋予卷积对输入数据的多维度敏感性，进而使其能够捕获更细腻的上下文信息。因此，ODConv 能显著增强卷积神经网络在特征提取方面的性能。

尽管 ODConv 展现出了卓越的性能，但其路由函数的设计仍侧重于考虑原始输入特征的通道信息，仅仅在维度变换后直接应用于卷积核的各个维度。这种处理策略虽然有效，但在深入挖掘适合于卷积核的注意力机制方面仍有提升的余地。

2.4 基于特征图的注意力机制

卷积核上的注意力机制使得网络能够依据不同的输入动态调节自身的权重参数，而特征图上的注意力机制则专注于突出网络感兴趣的特征区域。这两种机制均能有效提高卷积神经网络在特征提取方面的效率，从而优化视觉任务的处理效果。与基于卷积核的注意力机制相比，特征图上的注意力应用更少的参数和计算资源，更容易与现有网络架构相融合。此外，特征图层面的注意力机制，由于其作用于特征图而非卷积核，可以无缝集成进各类网络结构，而且不需要根据不同的卷积神经网络架构进行定制化修改，提供了一种更为通用的注意力增强手段。本节接下来将深入探讨几种代表性的基于特征图的注意力机制。

2.4.1 压缩和激励网络

压缩和激励网络 (Squeeze-and-Excitation Networks，SENet)[14]的出现，标志着在深度学习领域，尤其是计算机视觉领域的一项重大进展。其核心在于引入了 Squeeze-and-Excitation 模块，作用类似于人类的选择注意力机制，能够显式地建模特征通道之间的相互依赖性，从而自适应地调整特征在通道维度上的权重，选择出更为重要的特征通道。此方法使得 SENet 能够在不显著增加计算复杂度的前提下，增强网络结构的表征能力。

SE 模块的工作流程分为两个阶段，如图2-5所示。首先，通过 Squeeze 操作将全局空间信息聚合到一个通道描述符中，该过程通常使用全局平均池化实现，

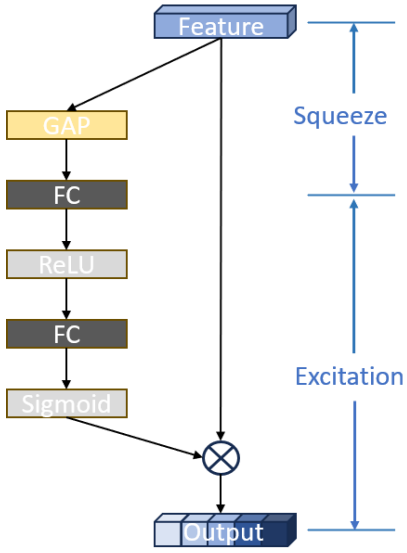


图 2-5 SENet 工作流程图

以确保模块能够捕捉到对重新校准通道特征至关重要的全局上下文信息。随后，在 Excitation 阶段，采用一种简单的门控机制 (通常使用两个全连接层和一个非线性激活层) 来捕获通道之间的依赖关系。Excitation 阶段的输出是一系列注意力系数。将这些注意力系数作用于原始特征图的通道维度，能够有效地强调重要特征通道并抑制次要特征通道。

尽管 SENet 展现出了较高的有效性并得到了广泛的应用，但并非没有缺陷。例如，在 Squeeze 模块中，全局平均池化过于简单，通常无法捕获复杂的全局信息；在 Excitation 模块中，两个全连接层增加了模型的复杂性，等等。所以虽然 SENet 在性能提升方面取得了显著成果，但其机制在一定程度上属于启发式，促进了注意力机制相关研究的蓬勃发展。

2.4.2 高效通道注意力机制

为了避免注意力机制给网络带来的复杂度过高，SENet 使用降维操作减少了通道数量。然而，这种降维策略无法直接对权重向量和输入特征之间的对应通道进行建模，从而降低了注意力系数的质量。为了克服这个缺点，Wang[16]提出了一种高效通道注意力机制 (Efficient Channel Attention Module，ECA)，它使用一维卷积代替降维操作来学习通道之间的交互信息。

如图2-6所示，ECANet 具有与 SENet 块相似的工作流程，包括用于聚合全

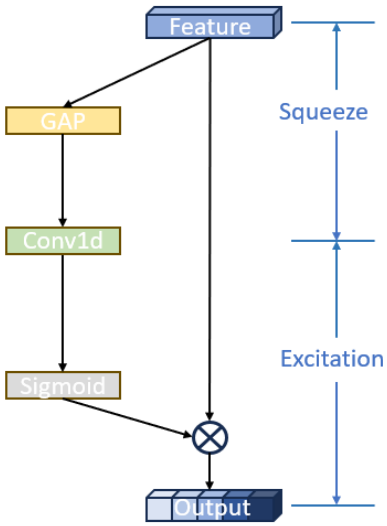


图 2-6 ECANet 工作流程图

局空间信息的 Squeeze 模块和用于建模跨通道交互的 Excitation 模块。不同的是，Excitation 模块不采用先降维再还原的全局间接对应方式，而是仅考虑每个通道与其 k 个最近邻通道之间的直接交互。从结构上来看，ECANet 通过使用通道数为 k 的一维卷积而非全连接层来捕捉通道间的相互依赖性，显著减少了模型的参数量和计算复杂度。参数 k 决定了通道间交互的范围，根据当前特征的通道数 C 确定其取值，如下式所示：

$$k = \psi(C) = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{\text{odd}}, \tag{2-12}$$

其中，γ 和 b 是超参数，$|x|_{\text{odd}}$ 表示距离 x 最近的奇数值。

ECANet 相比于 SENet，主要在计算效率、参数量、自适应性及实现难度等方面进行了显著的改进，使其成为一种更为高效和实用的通道注意力机制。这些改进使得 ECANet 能够在保持甚至提升性能的同时，减少模型的计算负担和资源消耗，特别适合于资源受限的环境。

2.4.3 卷积注意力模块

与 SENet 仅在通道维度上施加注意力不同，卷积注意力模块 (Convolutional Block Attention Module，CBAM)[15]通过顺序地关注通道和空间两个维度的关键特征，极大地提升了网络在提取和利用图像信息方面的能力。

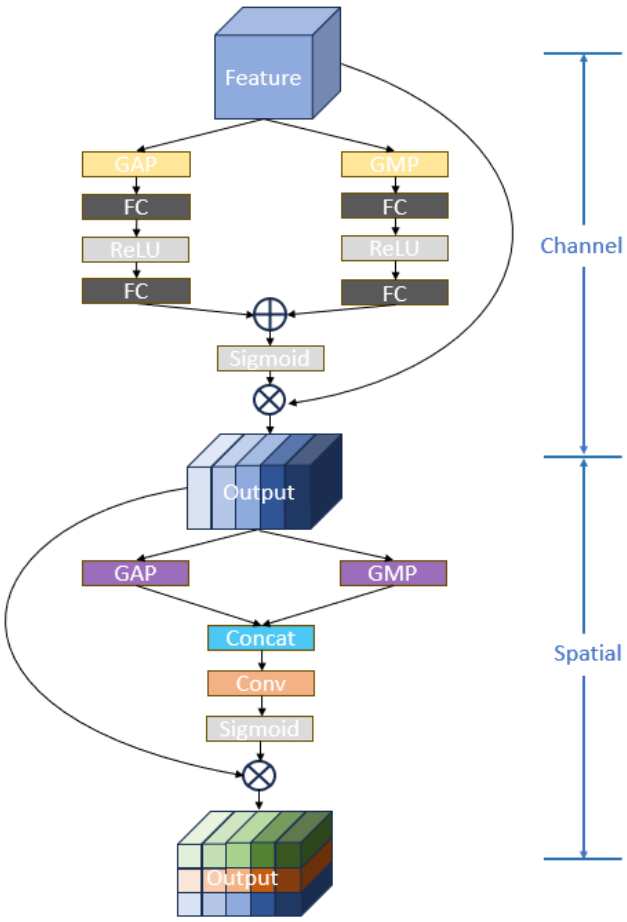


图 2-7 CBAM 工作流程图

如图2-7所示，CBAM 由两个连续的子模块组成：通道注意力模块和空间注意力模块，它们分别针对输入特征生成一维的通道注意力系数和二维的空间注意力系数。其中，通道注意力模块的设计方式与 SENet 类似，不同之处在于它不止采用一种类型的池化操作来聚合全局信息，而是同时使用全局平均池化和全局最大池化两个并行分支，通过这两种池化操作的融合，增强了对全局信息的聚合能力。空间注意力模块则压缩特征的通道信息，对其空间关系进行建模，从而丰富了特征在空间语义上的表征。与通道注意力机制不同，空间注意力模块采用较大的卷积核 (如 7×7) 来整合空间信息，以突出局部区域的语义。

CBAM 将通道注意力和空间注意力依次结合起来，有效地利用了特征的跨通道关系与空间关系。在通道维度上，CBAM 通过分析各个通道对任务的贡献度，自适应地调整通道权重，从而增强了对有利于任务执行的特征通道的响应。在空间维度上，CBAM 进一步细化了对图像局部区域的关注，通过空间注意力机制突出了图像中对任务目标有重大贡献的区域，这种对局部细节的强调使得

网络能够更加精准地捕获和利用图像信息。但由于 CBAM 直接对全局信息进行整合以生成注意力，在网络层次越来越深的情况下存在一定局限，使得高维特征图上的注意力系数缺乏一定准确性。

2.5 本章小结

本章主要介绍了通用视觉注意力机制的概念以及其作用对象的基本原理，随后依次介绍了基于卷积核的注意力以及基于特征图的注意力中几种具有重大影响力的方法，并且分析了不同方法的特点所在。

第三章　局部感知卷积核注意力机制

本章重点关注基于卷积核的注意力机制设计。首先，我们讨论了在卷积核中融入注意力机制的重要性，并分析了当前方法的不足之处和潜在的改进空间。在此基础上，我们提出了一种新颖的基于局部特征的卷积核注意力机制——LADConv。通过系列实验证明，LADConv 在各种视觉任务中的性能均优于现有的卷积核注意力方法，证明了其卓越的性能和应用潜力。

3.1　基于局部特征的卷积核注意力机制

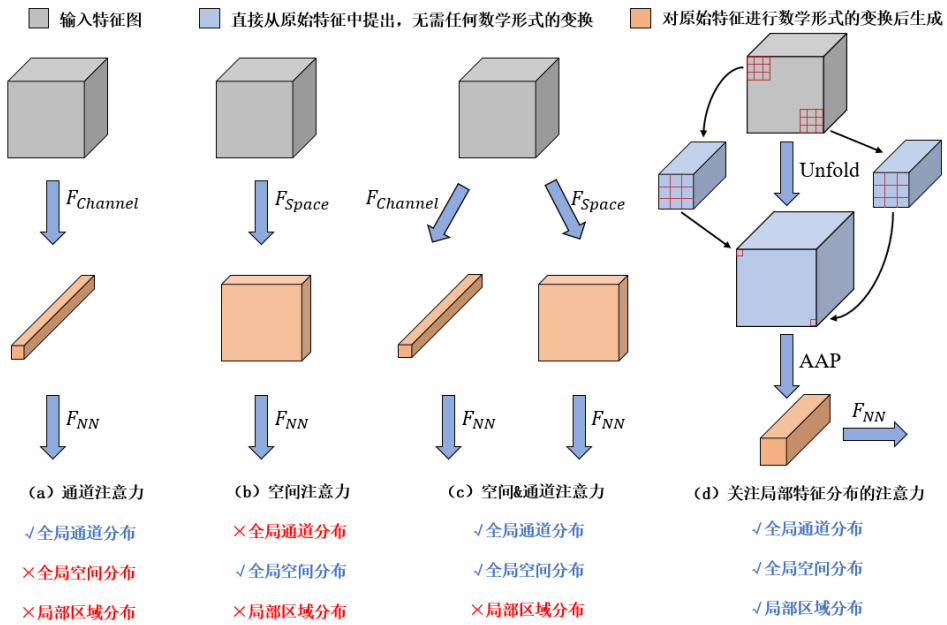


图 3-1　不同注意力机制的 Squeeze 模块对比示意图

在2.3节中，我们深入探讨了几种流行的基于卷积核的注意力机制及其特性。值得注意的是，这些方法大多直接采用了特征图注意力模型 SENet[14]中的 Squeeze and Excitation(SE) 模块来生成卷积核上的注意力系数。如图3-1所示，此类结构通常涉及对输入特征在空间维度的全局平均池化操作，使得生成的注意

力系数主要反映了特征在通道维度上的全局信息，从而忽视了局部区域和空间分布的重要性。鉴于卷积核在捕捉局部特征上的关键作用，SE 模块的局限性可能导致生成的注意力无法充分指导卷积核参数的动态调整。在空间维度上的注意机制中也出现了类似的例子，已有研究证明保留特征全局空间分布对于有效重塑特征在空间维度的权值至关重要[15,55]。因此，设计一种能够精准学习局部特征分布的注意力机制，对于增强卷积核捕获局部特征的能力显得尤为必要。

为了解决上述问题，本章提出了一种新型的基于局部特征的卷积核注意力机制 (Local-Aware Dynamic Convolution Attention，LADConv)。这种机制旨在通过生成更加合适的注意力系数来指导卷积核参数的动态调整，从而关注特征的局部区域。LADConv 的核心设计在于其中的 Squeeze 模块，如图3-1(d) 所示，该模块能够覆盖并捕捉卷积核所涉及的整个局部区域内的特征信息。通过有效地学习特征的多维分布，并特别强调对局部区域维度的关注，该模块显著提升了注意力系数的有效性。在经过 Squeeze 模块处理后，选取具有代表性的局部特征，并采用自注意力机制加强局部特征间的相互作用，以此生成注意力系数。值得注意的是，当卷积核的大小设置为 1 时，先前的卷积核注意力机制可以视为 LADConv 的一个特例。随着卷积核大小的增加，LADConv 展现了更为出色的局部特征捕捉能力。通过系列实验证实，LADConv 在各项任务中相比于传统的卷积核注意力机制展现了更为优异的性能。此外，我们还进行了广泛的消融实验，以凸显在核注意力设计中捕获局部特征的重要性。

接下来将详细介绍 LADConv 的设计理念及架构。首先，我们将概括性地阐述整体模型框架的构成，继而深入解析各个组成模块的结构与功能。

3.1.1 模型整体结构

如图3-2所示，LADConv 的结构主要由压缩 (Squeeze)，自注意力机制 (Self-attention) 和生成 (Generation) 三个部分的模块组成。压缩模块和自注意力模块负责提取并产生包含局部特征信息的描述符，而生成模块则负责将这些特征描述符转换成不同维度的注意力系数，并施加在卷积核的固定权重上。接下来，将对每个模块的细节进行详细讨论。

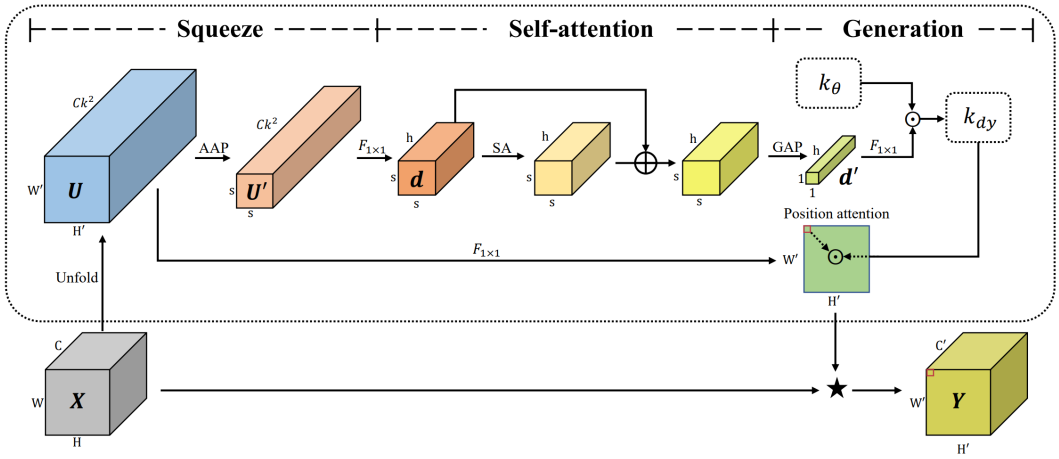


图 3-2 LADConv 模型结构图

3.1.2 特征压缩模块

在压缩模块的设计上，我们旨在生成能够精准反映输入特征局部分布信息的简单特征向量。并且，我们认为这些局部分布信息来源于卷积部核在原始输入特征图上所覆盖的局部区域。因此，在压缩模块中我们采用与卷积核滑动窗口一致的采样方式对这些局部分布信息进行采样，并且采样过程中保持一致的步长和填充策略。将一次卷积操作覆盖的局部区域的大小计为 $k \times k \times C$，为了简化后续的计算过程，我们将每个局部区域映射为一个 Ck^2 维的向量，并将这些向量串联起来，构建成一个新的特征图，记为 $U \in \mathbb{R}^{H' \times W' \times Ck^2}$。上述操作称为 “Unfold”，其形式化的表述如下：

$$\begin{aligned} U &= \text{Unfold}(X), \\ U_{ij} &= \left[X_{i-\lfloor k/2 \rfloor, j-\lfloor k/2 \rfloor}; \cdots; X_{i+\lfloor k/2 \rfloor, j+\lfloor k/2 \rfloor} \right], \end{aligned} \tag{3-1}$$

其中，索引 i 和 j 代表输入特征图中像素的排列顺序，而 U_{ij} 表示位于位置 (i, j) 的局部区域中所有元素的集合。需要强调的是，展开 (Unfold) 操作只是对输入元素进行位置上的重新排列，并没有进行任何形式的数学变换。紧接着，我们对 U 采用自适应平均池化 (Adaptive Average Pooling, AAP) 以提取具有代表性的局部分布：

$$U' = \text{AAP}(U), \tag{3-2}$$

经池化操作后，$U' \in \mathbb{R}^{s\times s\times Ck^2}$，超参数 s 表示沿高度和宽度维度所采集到的样本数。为了更好地捕获每个局部分布中的深层特征信息，在 U' 上使用 1×1 的卷积层进行降维：

$$d = \sigma\left(\mathrm{BN}\left(\mathrm{F}_{1\times 1}\left(U'\right)\right)\right), \tag{3-3}$$

其中，1×1 卷积的输出通道数 h 为 $\max(\lambda, C//r)$，λ 和 r 是超参数。此外，BN 和 σ 分别表示批量归一化和 GELU 激活函数[56]。

从直观上讲，我们设计的压缩模块能有效捕获输入数据内固有的局部特征属性，与卷积核的先验归纳偏好相契合。在3.2.6节中，通过对压缩模块的消融实验，我们进一步验证了这一论点。在这些实验中，我们探索了多种不同的压缩策略以生成特征描述符，其中基于局部特征提取的方法展示出了最显著的性能提升。

3.1.3 通过自注意力集成局部特征

将压缩模块的结果记为 d，其中包含了 s^2 个局部特征。值得注意的是，直接使用这些特征来生成注意力系数容易受到排列歧义的影响。一种直观的方法是对 s^2 个局部特征进行平均，然而这种方法忽视了局部特征的多样性，削弱了对关键局部细节的表征能力。为了克服这一挑战，我们采用自注意力 (SA) 机制来学习这些局部特征之间的相互依赖关系，从而保留了信息丰富的局部特征。具体而言，我们将 d 重塑为一个 $s^2\times h$ 的张量，并应用层归一化处理。然后，将 s^2 维度视为序列维度来应用自注意力机制。相关公式如下所示：

$$\begin{aligned} d_{Norm} &= \mathrm{LN}(d), \\ \mathrm{SA}(d) &= \mathrm{Softmax}(\frac{F_Q\left(d_{Norm}\right) F_K\left(d_{Norm}\right)^\top}{\sqrt{h}}) F_V\left(d_{Norm}\right), \\ d &:= \mathrm{SA}(d) + d, \end{aligned} \tag{3-4}$$

其中，$F_Q(\cdot)$，$F_K(\cdot)$ 和 $F_V(\cdot)$ 对应于输出尺寸为 h 的三个全连接层。为了公式的简洁，在公式3-4中省略了矩阵变形操作。最后，将描述符沿着 s^2 轴进行全局平均池化操作，形式化如下：

$$d' = \mathrm{GAP}(d), \tag{3-5}$$

形成一个 h 维的局部特征描述符 d' 用来生成注意力系数。

输入特征中的局部分布天然具有多样性，本节所提出的自注意力模块能够有效捕捉局部特征间的依赖关系。尽管如此，公式3-5中的平均池化操作仍然会导致这种多样性的丢失，但通过增加超参数 s，可以减小这种损失，为自注意力模块提供更多的局部细节信息。本章的3.2.6小节，展示了 s 的经验设置，实现了准确性与计算效率之间的平衡。

3.1.4 生成卷积核注意力机制

在将局部特征整合成紧凑向量 d' 之后，我们将重心转移到了如何在卷积核中施加注意力机制。如图3-3所示，注意力不仅施加在卷积核的内部空间中，也施加在卷积核移动时的空间位置上。本章在卷积核内部空间中施加注意力机制的方法与 ODConv[31] 中所使用的方法保持一致。为确保图示清晰易懂，我们在图中仅描述了核内部空间中的空间和输入通道维度，以及它们相应的注意力值。

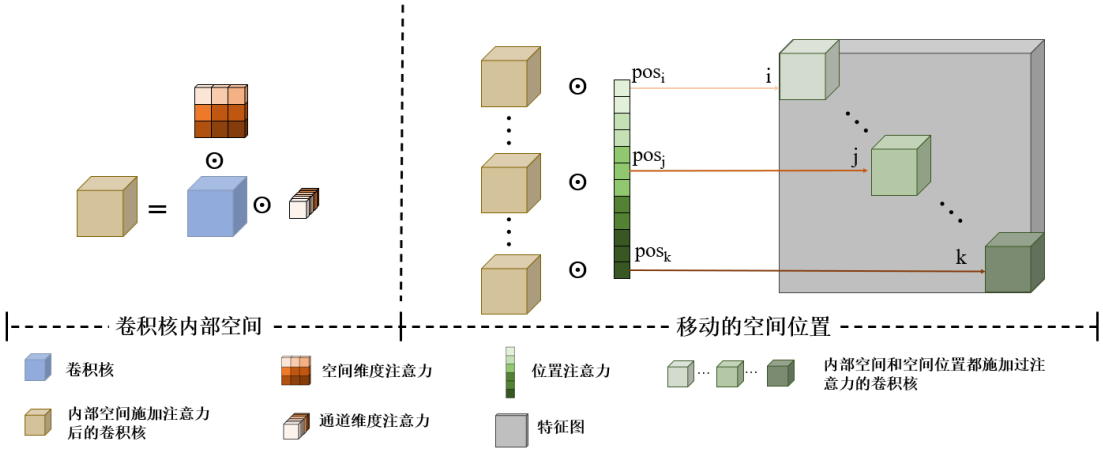


图 3-3 LADConv 的生成模块示意图

我们使用 A_{in}、A_{out}、A_{s} 和 A_{p} 分别表示输入通道维度、输出通道维度、卷积核空间维度和卷积核并行维度的注意力系数。每个系数都是通过下列方式产生的：

$$A = \text{Act}\left(F_{1\times1}\left(d'\right)\right), \tag{3-6}$$

其中，$\text{Act}(\cdot)$ 表示激活函数，对于 A_{in}、A_{out} 和 A_{s}，我们使用 Sigmoid 函数作为激活函数，而对于 A_{p}，激活函数为 Softmax 函数。当提供了 n 个静态卷积核 $k_\theta \in \mathbb{R}^{n\times k\times k\times C\times C'}$ 时，通过将注意力系数与静态卷积核的相应维度相乘，并随

后在并行卷积核维度上进行聚合来形成参数可动态调整的卷积核。这一过程可表示为如下形式：

$$\begin{aligned} k_{dy} &= A_{\mathrm{p}} \odot A_{\mathrm{s}} \odot A_{\mathrm{in}} \odot A_{\mathrm{out}} \odot k_{\theta}, \\ k_{dy} &:= \mathrm{Sum}_{p}\left(k_{dy}\right), \end{aligned} \tag{3-7}$$

$\mathrm{Sum}_{p}(\cdot)$ 代表了在并行卷积核维度上，将多个施加注意力系数后的静态卷积核聚合为一个动态卷积核的过程。鉴于卷积核在特征图上滑动时会以相同的参数遍历不同空间位置，我们引入了位置注意力机制以提升卷积核的动态调整能力，使卷积核在同一特征图的不同位置上也能根据实际需要灵活调整参数。需要注意的是，由于局部特征描述符 d' 缺乏细粒度空间分布，从而无法用于生成位置注意力系数。因此，我们从更原始的表示中提取位置特征。回想一下，U 中的每个像素代表了某个卷积核一次移动所覆盖的局部区域。因此，我们在 U 上使用 1×1 卷积来生成位置注意力系数，可以形式化地表示为：

$$P_{ij} = \mathrm{Sigmoid}\left(F_{1\times 1}\left(U_{ij}\right)\right), \tag{3-8}$$

最后，(i, j) 位置的输出特征值可以表示为：

$$Y_{ij} = \mathrm{Sum}\left(\left(P_{ij} \cdot k_{dy}\right) \odot U_{ij}\right), \tag{3-9}$$

将公式3-9中的 k_{dy} 展开，等价于矩阵 P 与卷积结果 $k_{dy} \star X$ 沿着空间维度的乘积。值得一提的是，我们的位置注意力机制与传统的空间注意力方法不同。在传统的空间维度注意力机制中，(i, j) 位置的注意力是通过使用一个较大尺寸的卷积核 (默认是 7×7) 来捕获局部交互生成的。与之相反，我们将卷积核在 (i, j) 位置所覆盖的区域直接映射到相应的位置注意力上，其中局部大小与卷积操作的卷积核大小保持一致。这一要求确保了其他像素不会干扰当前局部区域内卷积核捕获的特征。我们通过3.2.6节中的消融实验表明，使用不一致的局部大小生成位置注意力会导致性能下降，这强调了匹配局部大小的重要性。

3.1.5 LADConv 的扩展形式

1×1 卷积

LADConv 的压缩模块能够捕获与卷积核大小 k 相匹配的局部区域特征。然而在卷积神经网络中存在很多 k 为 1 的卷积核，此时局部区域特征变得过于简化，丢失了关键的空间模式，展开后的特征图 U 等价于原始输入 X。此外，1×1 卷积通常用于高维特征组合，在这种情况下原始的 LADConv 可能会给网络增加相当大的计算负担。因此，我们通过几种方式简化了 1×1 卷积核上的 LADConv。首先，将展开操作变为恒等变换，并将自适应平均池化 (AAP) 的样本数量 s 设置为 1(即全局平均池化 GAP)。接下来，将自注意力机制学习模块替换为一个简单的全连接层进行维度变换。最后，在注意力生成模块中，仅在 1×1 卷积核的输入和输出通道维度上施加注意力机制。

由于 1×1 卷积核所提取出的局部特征本质上等价于原始特征图上每个像素点的整条通道特征，经过上述操作，1×1 卷积核上的压缩模块将被简化为先前的卷积核注意力压缩模块的形式。换句话说，以往的工作可以近似为将 1×1 卷积核的局部特征直接作用在大小为 $k \times k$ 的卷积核上 ($k > 1$)，这将导致产生的局部特征描述符存在一定局限性。

深度可分离卷积

深度可分离卷积 (Depthwise Separable Convolution) 是一种特殊的卷积操作，广泛应用于轻量级或计算效率要求高的深度学习模型中[57-59]。深度可分离卷积的核心思想是将一个完整的卷积运算分解为两步进行，即深度卷积 (Depthwise Convolution，DWConv) 与逐点卷积 (Pointwise Convolution，PWConv)。其中 PWConv 实则就是 1×1 的卷积层，而 DWConv 则是对输入特征的每个通道单独进行卷积计算，为了在 DWConv 上施加合适的注意力机制，我们将 LADConv 中的各种设计扩展到逐通道级别。

首先，逐通道级别的展开操作结果 $U \in \mathbb{R}^{H' \times W' \times C \times k^2}$ 可以被表示为如下形式：

$$U_{ijl} = \left[X_{i-\lfloor k/2 \rfloor, j-\lfloor k/2 \rfloor, l}; \cdots; X_{i+\lfloor k/2 \rfloor, j+\lfloor k/2 \rfloor, l} \right], \tag{3-10}$$

其中，l 表示输入特征的通道次序。自适应平均池化操作 (AAP) 作用于 $H' \times W'$ 维度，生成 C 个尺寸为 $s \times s \times k^2$ 的特征图。我们对每个特征图采用权重共享的方式应用 1×1 卷积和自注意力机制融合局部特征。这一过程将产生 C 个局部特征描述符，形成一个 $C \times h$ 的张量。然后，我们在局部特征描述符之间应用额外的自注意力进一步学习有代表性的局部特征。在注意力生成模块中，注意力独立地应用于卷积核的每个输入通道。虽然移除了输出通道的注意力系数 A_{out}，但是通过逐通道全连接层生成了剩余维度的注意力系数 A_{in}，A_{s} 和 A_{p}。

值得注意的是，原始的 LADConv 实际上不作任何修改也能直接作用于 DW-Conv 上，类似于先前的卷积核注意力机制处理方式。然而，这种方法忽略了 DWConv 的固有通道独立性。通过3.2.3小节中的实验可以证明，我们为 DWConv 专门定制的修改版 LADConv 有效地解决了这一局限性。基于这一策略，LAD-Conv 可以无缝集成到各种分组卷积结构中。

3.2 实验与分析

3.2.1 实验数据分析

为了验证LADConv在实际视觉任务中的表现,本文分别在ILSVRC2012(ImageNet Large Scale Visual Recognition Challenge 2012)[45]、CIFAR-100(Canadian Institute for Advanced Research, 100 classes)[44]、MS COCO 2017(Microsoft Common Objects in Context)[46]以及 PASCAL VOC 2007+2012(The PASCAL Visual Object Classes)[60]四个常用的视觉任务数据集上进行训练和测试，其中 ImageNet 和 CIFAR100 数据集用于图像分类，MS COCO 和 PASCAL VOC 用于目标检测任务，具体数据样本划分如表3-1所示。

表 3-1 各个数据集的样本划分

数据集	训练集 (张)	测试集 (张)
ILSVRC2012	1 281 167	50 000
CIFAR-100	50 000	10 000
MS COCO	118 287	5 000
PASCAL VOC	8 216	8 333

ILSVRC2012

ILSVRC2012 数据集是 ImageNet 数据集的一个子集 (别名 ImageNet1K)，提供了一个复杂多样的图像分类测试场景，包含了约 128 万张人工标注的训练集图片及 5 万张验证集图片。这些图片总共被分为 1 000 个类别，覆盖了从日常用品到各类动植物的广泛范畴。ILSVRC2012 数据集的特点在于其规模和多样性，它不仅包含了大量的图像数据，而且这些数据来源多样，包括网络图片和专业拍摄的照片等，保证了图片在内容、背景、光照条件等方面的多样性和真实性，对图像分类任务的研究工作产生了深远的影响。

CIFAR-100

与规模较大的 ImageNet 数据集不同，CIFAR-100 数据集中包含的图片数量较少，尺寸也相对较小。CIFAR-100 数据集是 Tiny Images 数据集的一个子集，由 60 000 张尺寸为 32×32 的彩色图像组成。这些图片被细分为 100 个不同的类别，进一步归属于 20 个超类。每个类别包含 500 张训练图片和 100 张测试图片，提供了一个具有挑战性的测试环境，用于开发和评估图像识别与分类算法。尽管图片分辨率较低，CIFAR-100 数据集的类别细分和多样性仍然使其成为测试深度学习模型和其他计算机视觉技术的理想选择之一。

MS COCO

COCO 数据集是计算机视觉研究中一个常用的数据集，它提供了超过 12 万张的训练图片和 5 万张验证图片，覆盖 80 个对象类别，支持目标检测、实例分割和图像标注等多种任务。COCO 数据集因图像内容的广泛性、精确的对象轮廓分割以及详尽的图像注释而闻名，其背景涵盖了日常生活中的多种环境，如海滩、城市街道和乡村田野等。在 COCO 数据集中，几乎所有类别的注释数量都超过了 1 000 个实例。COCO 数据集根据对象的尺寸将其分为大尺寸对象 (尺寸超 96×96 像素)、中等尺寸对象 (尺寸在 32×32 到 96×96 像素之间) 和小尺寸对象 (尺寸小于 32×32 像素)，分别占据了数据集的 24%、32% 和 41%。这一分类表明，在目标检测任务中，COCO 数据集提出了相当难的挑战。

PASCAL VOC

同样，我们在目标检测任务中也选取了一个规模较小的数据集进一步验证 LADConv 的性能。为了使数据样本尽可能充分，我们将较为流行的 PASCAL VOC 2007 版本和 2012 版本合并使用，共计约 20 000 张图像，40 000 个目标。它包含 20 个不同的对象类别，涵盖了人、各种动物、交通工具和日常物品等，提供了详细的标注信息，包括类别标签和对象边界框。虽然其规模不如一些更大的数据集，如 ImageNet 或 COCO，但在研究早期，PASCAL VOC 以其高质量的标注和对多种计算机视觉任务的支持，在视觉算法的评估和比较中仍然扮演着重要角色，特别是在目标检测和图像分割方面。

3.2.2 评价指标

图像分类

对于图像分类任务而言，在 ILSVRC2012 数据集上进行的验证实验采用了 Top-1 和 Top-5 两个常用的评估分类模型性能的指标，而为了方便简洁，在 CIFAR-100 数据集上进行的验证实验仅采用了 Top-1 指标。

Top-1 准确率是最直接的性能指标。对于给定的输入 (如一张图片)，模型会预测一个最有可能的类别作为输出。如果这个预测的类别与真实类别 (即标签) 相匹配，则认为这次预测是正确的。形式化地说，给定一个样本集合 S，对于每个样本 $s_i \in S$，模型产生的预测为 p_i(预测的最可能类别)，而真实的类别为 g_i。则 Top-1 准确率定义为预测正确的样本数量占总样本数量的比例：

$$\text{Top-1 Accuracy} = \frac{1}{|S|}\sum_{i=1}^{|S|} I\left(p_i = g_i\right), \tag{3-11}$$

其中，I 是指示函数，如果 $p_i = g_i$，则其值为 1，否则为 0。

Top-5 准确率提供了模型性能的另一个视角，特别是当分类任务中的类别非常多时。在这种情况下，即使模型的最高置信度预测 (Top-1) 是错误的，模型仍可能在其前 5 个最高置信度的预测中包含了正确的类别。对于每个样本 s_i，模型产生一组预测 $P_i = \{p_{i1}, p_{i2}, p_{i3}, p_{i4}, p_{i5}\}$，这组预测包含了模型认为最可能的 5 个类别。如果真实类别 g_i 出现在这个预测集合 P_i 中，那么这次预测被认为是

正确的。因此，Top-5 准确率可以定义为：

$$\text{Top-5 Accuracy} = \frac{1}{|S|}\sum_{i=1}^{|S|} I\left(g_i \in P_i\right), \tag{3-12}$$

Top-1 和 Top-5 准确率都是衡量模型在多类别的分类任务中性能的重要指标，Top-5 准确率特别适用于类别众多且相互之间较为相似的情况。

目标检测

对于目标检测任务而言，本章使用最通用的评价指标 mAP(mean Average Precision) 来评价 LADConv 的表现。为了全面理解 mAP，我们首先需要定义召回率 (Recall)、精确度 (Precision) 和 IoU(Intersection over Union) 的概念及其形式化表达。

表 3-2 TP、FP、FN 和 TN 的定义

	标签为真	标签为假
预测为真	True Positive(TP)	False Positive(FP)
预测为假	False Negative(FN)	True Negative(TN)

如表3-2所示，在说明精确度和召回率之前，需要对更基础的概念 TP(True Positive)、FP(False Positive)、FN(False Negative) 和 TN(True Negative) 有一定了解。精确度衡量的是模型正确识别的正样本占模型识别为正样本的总数的比例。形式化地表示为：

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}, \tag{3-13}$$

其中，TP 是正确识别的正样本数量，FP 是错误识别为正样本的负样本数量。对应到目标检测任务中，TP 即样本预测类别与真实类别一致，并且样本预测框的 IoU 大于阈值 t 的样本数。召回率衡量的是模型正确识别的正样本占所有实际正样本的比例。形式化地表示为：

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}, \tag{3-14}$$

FN 是未检测到的正样本数量。IoU 是评估目标检测中预测边界框准确度的指标，

计算为预测框与真实框交集的面积与它们并集的面积之比：

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}, \tag{3-15}$$

通常，IoU 阈值被设定来判断预测框是否准确。例如，在某些任务中，IoU > 0.5 的预测可能被认为是正确的。

mAP 是所有类别平均精确度 (AP) 的平均值，其中 AP 是给定类别在不同召回率水平下最高精确度的平均值。对于每个类别，我们首先根据模型预测的置信度对样本进行排序，然后计算每个可能的召回率水平的精确度，并取这些精确度的最大值 (以优化模型性能)。AP 是这些最大精确度的平均值。最后，mAP 是所有类别 AP 值的平均值：

$$\text{mAP} = \frac{1}{C}\sum_{c=1}^{C} AP_c, \tag{3-16}$$

在目标检测任务中，AP 通常在多个 IoU 阈值下计算，以充分考虑预测框与真实框的匹配程度。例如，mAP@.5 是在 IoU 阈值为 0.5 时计算的 mAP，而 mAP@[.5:.95] 是在 IoU 从 0.5 到 0.95(以 0.05 为步长) 的范围内计算的 mAP 值的平均值，一般直接称 mAP@[.5:.95] 为 mAP。

3.2.3 ImageNet 对比实验

实验设置

我们提供了两种用于比较的 LADConv 模型：LADConv(1×) 和 LADConv(4×)，其中 ($n\times$) 表示模型使用了 n 个并行的静态核。这些模型的主干架构包括 ResNet-18、ResNet-50、ResNet-101[8]、MobileNetv2[42]以及 ConvNeXt-Tiny[43](简称 ConvNeXt-T)。

在 ImageNet 数据集上的实验部分，我们深入地对 LADConv 及其相关方法进行了细致比较，旨在凸显 LADConv 的优越性能。针对 ResNet-18、ResNet-50 和 MobileNetV2 这三种网络架构，我们选取了 CondConv[30]与 DyConv[52]这两种著名的卷积核注意力方法；以及 DCD[54]与 SD-Conv[61]，这两种更为轻量级的卷积

核注意力方案；此外还与 ODConv[31]这种表现卓越的卷积核注意力技术进行了系统比较。特别是在 ResNet-18 和 ResNet-50 的架构下，我们对包括上述技术在内的多种方法进行了广泛的评估。对于 ResNet-101 的架构，我们主要与 ODConv 进行了对比，因为其他几种技术在 ResNet-101 上尚未有公开的结果和实现细节。需要指出的是，在 ConvNeXt 架构的实验中，目前尚缺乏专为该架构设计的、公开可用的卷积核注意力机制的实现版本。

为了确保比较的公平性，我们保持了与这些对比方法一致的训练设置。在 ResNet 的实验中，模型通过随机梯度下降 (stochastic gradient descent，SGD)[62]训练了 100 个周期，其中权重衰减系数为 10^{-4}，动量 (Momentum) 为 0.9，初始学习率设定为 0.1，并且每 30 个周期降低为原来的十分之一。对于 MobileNetV2 的实验，训练周期延长到 150 个，权重衰减系数为 4×10^{-5}，初始学习率设为 0.05，并遵循单个余弦周期内递减至零的设置。为了排除干扰并与之前的工作保持一致，我们也在前 10 个周期采用了温度退火 (temperature annealing)[63]训练策略，并对 ResNet-18 应用了 0.1 的 dropout 率，对其他模型应用了 0.2 的 dropout 率。ConvNeXt 实验中的训练配置与该模型论文中的配置一致。

LADConv 模型还包含几个预设置的超参数。在普通的卷积结构中，降维阈值 λ 和 r 都被设置为 16。对于 MobileNetV2 中的 DWConv 结构，降维剩余通道数量 h 设置为 4，而对于 ConvNeXt-T 中的 DWConv 结构，h 设置为 8。在 ResNets 和 MobileNetV2 的最后一层中，我们将池化后样本数 s 设置为 3，而在其余层次中，s 被设置为 5。在 ConvNeXt-T 中，我们特别将 7×7 的卷积替换为配置了 s 为 2 的 LADConv 结构。

ResNet 实验结果

基于 ResNet 模型的验证集准确率、模型计算力 (FLOPs) 以及参数量如表格3-3所示。我们不难发现，传统的卷积核注意力 CondConv 和 DyConv 通过引入额外的静态卷积核参数的方式增强了基准模型的性能，但也导致模型总参数量显著增加 (对于 ResNet-18 分别增加了 7.0/3.9 倍，对于 ResNet-50 分别增加了 5.0/3.9 倍)。而我们的 LADConv(1×) 与基准模型保持相当的参数量和 FLOPs，性能却超过了除 ODConv(4×) 以外的所有卷积核注意力方法。当同样使用多个静态卷积核训练 LADConv 时，我们的 LADConv(4×) 达到了最优的结果，以极小

表 3-3 LADConv 在 ImageNet 验证集上基于 ResNet 系列模型的测试结果

模型	Top-1 准确率 (%)	Top-5 准确率 (%)	算力 (GMac)	参数
ResNet-18	70.25	89.38	1.81	11.69M
+CondConv(8×)	71.99(↑1.74)	90.27(↑0.98)	1.89	81.35M
+DyConv(4×)	72.76(↑2.51)	90.79(↑1.41)	1.86	45.47M
+DCD	72.33(↑2.08)	90.65(↑1.27)	1.84	14.70M
+SD-Conv	73.30(↑3.05)	-	1.51	23.20M
+ODConv(1×)	73.10(↑2.85)	91.10(↑1.72)	1.84	11.94M
+ODConv(4×)	73.97(↑3.72)	91.35(↑1.97)	1.92	44.90M
+LADConv(1×)	73.55(↑3.30)	91.19(↑1.82)	1.85	12.63M
+LADConv(4×)	**74.38(↑4.13)**	**91.52(↑2.14)**	1.93	45.58M
ResNet-50	76.23	93.01	3.86	25.56M
+CondConv(8×)	76.70(↑0.47)	93.12(↑0.11)	3.98	129.86M
+DyConv(4×)	76.82(↑0.59)	93.16(↑0.15)	3.97	100.88M
+DCD	76.92(↑0.69)	93.46(↑0.45)	3.94	29.84M
+SD-Conv	77.40(↑1.17)	-	3.40	54.00M
+ODConv(1×)	77.96(↑1.73)	93.84(↑0.83)	3.92	28.64M
+ODConv(4×)	78.52(↑2.29)	94.01(↑1.00)	4.08	90.67M
+LADConv(1×)	78.28(↑2.05)	93.90(↑0.89)	3.92	29.37M
+LADConv(4×)	**78.86(↑2.63)**	**94.06(↑1.05)**	4.08	91.41M
ResNet-101	77.41	93.67	7.57	44.55M
+ODConv(1×)	78.98(↑1.57)	94.38(↑0.83)	7.68	50.82M
+ODConv(2×)	79.27(↑1.86)	94.47(↑0.80)	7.80	90.44M
+LADConv(1×)	**79.31(↑1.90)**	**94.49(↑0.82)**	7.69	52.17M

的额外计算成本在 ResNet-18/ResNet-50 上分别超过 ODConv(4×)0.41%/0.34% 的精度。在 ResNet-101 的比较中，LADConv(1×) 在比 ODConv(2×) 少 42.32% 的参数前提下展现出更佳的优化能力。

在表3-3中，我们通过与其他卷积核注意力方法的对比，展示了 LADConv(1×) 和 (4×) 模型在 ResNet-18、ResNet-50 以及 ResNet-101 上的出色性能。值得注意的是，在 ResNet-18 上 LADConv 取得了 4.13% 的准确率显著提升，相较于 ResNet-50 上的 2.63% 提升更为突出。我们认为这种性能差异源于网络结构的不同。一方面，我们的卷积核注意力机制旨在增强核的局部特征提取能力，在大于 1×1 的卷积核上施加的注意力效果更为明显。因此，对于仅采用 3×3 卷积的 ResNet-18，卷积核注意力机制能够带来更优的性能提升。另一方面，由于 ResNet-50 在每个单元中都采用 1×1 卷积以实现特征的高维通道扩展，单纯依赖于增强局部特征提取的卷积核注意力可能不足以充分发挥作用。这一观察为第 4 章探索特征图级别的注意力机制——这种聚焦于突出丰富信息通道的特征重校准方法，提供了新的视角和启示。

MobileNetV2 实验结果

表 3-4 LADConv 在 ImageNet 验证集上基于 MobileNetV2 系列模型的测试结果

模型	Top-1 准确率 (%)	Top-5 准确率 (%)	算力 (MMac)	参数
MobileNetV2(1.0×)	71.65	90.22	300.8	3.05M
+CondConv(8×)	74.13(↑2.48)	91.67(↑1.45)	318.1	22.88M
+DyConv(4×)	74.94(↑3.29)	91.83(↑1.61)	317.1	12.40M
+DCD	74.18(↑2.53)	91.72(↑1.50)	318.4	5.72M
+SD-Conv	75.30(↑3.65)	-	261.9	7.7M
+ODConv(1×)	74.84(↑3.19)	92.13(↑1.91)	311.8	4.94M
+ODConv(4×)	75.42(↑3.77)	92.18(↑1.96)	327.1	11.52M
+LADConv*(1×)	75.20(↑3.55)	92.14(↑1.92)	378.0	7.37M
+LADConv(1×)	75.33(↑3.68)	92.16(↑1.94)	353.1	4.76M
+LADConv(4×)	**75.77(↑4.12)**	**92.25(↑2.03)**	367.8	11.33M
MobileNetV2(0.75×)	69.18	88.82	209.1	2.64M
+CondConv(8×)	71.79(↑2.61)	90.17(↑1.35)	223.9	17.51M
+DyConv(4×)	72.75(↑3.57)	90.93(↑2.11)	220.1	7.95M
+DCD	71.92(↑2.74)	90.20(↑1.38)	222.9	4.08M
+SD-Conv	73.20(↑4.02)	-	171.8	5.0M
+ODConv(1×)	72.43(↑3.25)	90.82(↑2.00)	217.1	3.51M
+ODConv(4×)	73.81(↑4.63)	91.33(↑2.51)	226.3	7.50M
+LADConv(1×)	73.70(↑4.52)	91.03(↑2.21)	252.5	3.48M
+LADConv(4×)	**74.08(↑4.90)**	**91.39(↑2.57)**	261.3	7.47M

在 MobileNetV2 架构上进行的实验进一步验证了 LADConv 在轻量级网络结构中的应用潜力，相关的比较结果已展示在表3-4中。我们发现，以往的卷积核注意力技术，如 CondConv(8×)、DyConv(4×) 以及 ODConv(4×)，虽然在性能上相较于基准模型有所提升，但却大量增加了模型的参数量，这与 MobileNetV2 旨在实现的轻量级设计初衷不符。与此形成鲜明对比的是，LADConv(1×) 在几乎不增加额外参数的前提下，就能实现与 ODConv(4×) 相媲美的性能，具体而言，在 MobileNetV2(1.0×) 和 MobileNetV2(0.75×) 上与 ODConv(4×) 的性能差距仅为 0.09% 和 0.11%。更值得一提的是，LADConv(4×) 版本在性能上更是达到了最好水平，相较于 ODConv(4×)，在 MobileNetV2(1.0×) 和 MobileNetV2(0.75×) 上分别实现了 0.35% 和 0.27% 的显著提升。

值得注意的是，在 MobileNetV2 上施加的 LADConv 卷积核注意力机制实际上采用了第 3.2.5.2 节所述的深度可分离变体形式。此外，我们还对 MobileNetV2 进行了原始 LADConv 结构的测试，这一结构在文中被称作 “LADConv*” 。该方法并未考虑到深度可分离卷积 (DWConv) 核心的通道独立特性，这与 ODConv 等传统卷积核注意力机制的处理方式类似。根据表3-4中的实验结果，我们观察到深度可分离的 LADConv 版本增加的额外参数量更少，同时在性能上也略有优

势，超过了原版 LADConv 结构。

ConvNeXt 实验结果

表 3-5 LADConv 在 ImageNet 验证集上基于 ConvNeXt-T 模型的测试结果

模型	Top-1 准确率 (%)	Top-5 准确率 (%)	算力 (GMac)	参数
ConvNext-T	82.05	95.86	4.48	28.59M
+LADConv(1×)	**82.40(↑0.35)**	**96.03(↑0.17)**	4.66	31.35M

在将 LADConv 卷积核注意力施加于 ConvNeXt-T 框架中时，我们遵循了 ConvNeXt 的训练准则，实施了为期 300 个周期的训练计划，并采取了多种数据增强策略。这些训练策略旨在协助模型达成高准确率，以逼近由模型能力界定的性能极限。值得注意的是，正如表3-5所展示的，我们在仅引入极少额外参数的前提下，实现了 0.35% 的显著准确率提升。这一结果表明，LADConv 能显著提升 ConvNeXt 的模型潜力。此外，ConvNeXt 采用了较大的核尺寸 ($k = 7$)，使得局部区域包含了更丰富的信息。我们认为，LADConv 捕获局部特征的能力在 ConvNeXt 的背景下被进一步凸显。

3.2.4 COCO 对比实验

实验设置

在目标检测任务中，我们选取了主流的 Faster R-CNN 框架[64]并采用特征金字塔网络 (Feature Pyramid Networks, FPN)[65]作为连接部分。鉴于 LADConv 具备良好的通用性，我们直接将 Faster R-CNN 的骨干网络替换为已经在 ImageNet 上完成预训练的 LADConv 模型 (对 ResNet50 和 MobileNetV2 均采用 LADConv(4×) 版本)，随后在训练数据集上对整体模型进行了为期 12 个周期的微调。该系列实验的实施借助了 MMDetection 工具箱[66]。

为了确保比较的公平性，我们在特征金字塔网络 (FPN) 的连接部分卷积层中并未引入任何注意力机制，并且对于采用 CondConv、DyConv、ODConv 以及 LADConv 构建的所有预训练模型，我们保持一致的数据预处理操作和超参数配置。在 MS-COCO 训练集上，这些检测器根据 1× 学习率方案进行微调，整个过程持续 12 个训练周期，其中在第 8 个和第 11 个周期时，我们将学习率分别降低

为之前的十分之一。遵循 DyConv 的实验设置，我们在 MS-COCO 数据集的实验中并未使用温度退火策略，以免影响模型性能。在验证过程中，IoU 阈值范围为从 0.5 到 0.95，以 0.05 为增量。

表 3-6 LADConv 在 COCO 验证集上基于 Faster R-CNN 模型的测试结果

骨干网络	mAP(%)	算力 (GMac)	参数
ResNet50	37.2	207.07(76.50)	43.80M(23.51M)
+CondConv(8×)	38.1	207.08(76.51)	133.75M(113.46M)
+DyConv(4×)	38.3	207.23(76.66)	119.12M(98.83M)
+DCD	38.1	207.20(76.63)	48.08M(27.79M)
+ODConv(1×)	39.0	207.18(76.61)	46.88M (26.59M)
+ODConv(4×)	39.2	207.42(76.85)	108.91M (88.62M)
+LADConv(4×)	**39.4**	207.48(76.91)	109.44M (89.15M)
MobileNetV2	31.3	122.58(24.45)	21.13M (2.22M)
+CondConv(8×)	33.7	122.59(24.46)	31.54M (12.63M)
+DyConv(4×)	34.5	123.01(24.88)	30.02M (11.12M)
+DCD	33.3	123.01(24.88)	23.34M (4.44M)
+ODConv(1×)	34.3	123.00(24.87)	22.56M (3.66M)
+ODConv(4×)	35.1	123.02(24.89)	29.14M (10.24M)
+LADConv(4×)	**35.4**	123.06(24.93)	28.92M (10.02M)

表格3-6展示了各种卷积核注意力方法在 MS-COCO 2017 验证集上的结果比较。关于 Flops 和 Params，括号中的数字是针对预训练的骨干模型 (不包括最后一个全连接层) 的，而其他数字则针对整个目标检测框架。从上表我们可以得出两个结论。首先，目标检测的性能与骨干网络的性能之间存在密切的关系。与基准模型相比，LADConv 模型将平均精度 (mAP) 提高了 2.2%/4.0%。其次，卷积核注意力在目标检测任务中与图像分类任务中带来的性能提升相似，在 ResNet-50 和 MobileNetV2 骨干网络上，LADConv 均取得了最佳结果。

3.2.5 小数据集上的对比实验

CIFAR-100

表 3-7 LADConv 在 CIFAR-100 验证集上基于各种模型的测试结果

模型	Top-1 准确率 (%)	模型	Top-1 准确率 (%)
MobileNetv1	65.98	VGG-16	78.58
MobileNetv1*	**67.91**	VGG-16*	**79.96**
DenseNet-121	77.01	ResNeXt-50	77.77
DenseNet-121*	**79.24**	ResNeXt-50*	**78.47**

在 CIFAR-100 数据集的实验中，我们将 LADConv(1×) 卷积核注意力集成到

不同的网络架构中，并在表3-7中用带有 $*$ 的模型表示，包括 VGGNets[6](具体为 VGG-16)、MobileNetv1[57]、ResNeXt(2x64d)[67]和 DenseNet(具体为 DenseNet-121)[68]。所有模型采用随机梯度下降进行训练，共 180 个周期，批量大小设置为 64，权重衰减设定为 0.0005，动量设定为 0.9。初始学习率设定为 0.1，每 60 个周期学习率除以 5。此外，为了确保一致性并避免随机初始化的影响，我们将随机种子固定在 1029，并对每组模型都进行了三次实验。

表3-7清晰地展示了 LADConv 能够显著提升所有参与测试的卷积神经网络模型的性能。这些实验毫无疑问地证实了 LADConv 在增强卷积神经网络性能方面的泛化能力。此外，值得注意的是，在 CIFAR-100 数据集上观察到的平均提升略小于在 ImageNet 上的提升，这种差异可能归因于潜在的过拟合问题。

PASCAL VOC

表 3-8 LADConv 在 PASCAL VOC 2007+2012 验证集上的测试结果

检测框架	骨干网络	mAP
Faster R-CNN	ResNet-50	79.5
	ResNet-50 (LADConv)	**80.9**
RetinaNet	ResNet-50	77.3
	ResNet-50 (LADConv)	**78.6**

同样，对于目标检测任务我们也在小规模数据集 PASCAL VOC 2007+2012 上进一步评估了模型的泛化能力。我们使用 ResNet-50+LADConv(1×) 作为骨干网络来捕获特征。该实验中采用的检测框架包括 Faster R-CNN 和 Retinanet[69]。如表3-8所示，LADConv 在 Faster R-CNN 和 RetinaNet 框架上分别比基准模型高出 1.4% 和 1.3%。基于我们的各种实验，我们相信 LADConv 能够在广泛的视觉任务和网络结构上展现其优越性。

3.2.6 消融实验

在本节中，我们进行了一系列的消融实验，旨在探究以下几个方面的问题：(1)LADConv 模型结构中每个组成部分的有效性；(2)LADConv 模型结构中超参数的经验设置；(3) 以及压缩模块中局部尺寸的影响。

压缩模块的对比

LADConv 中的压缩模块主要目的是构建卷积核感受域内的局部特征与其所对应的注意力系数之间的桥梁。为了评估该模块的有效性，我们选取了几种在现有研究中广泛使用的压缩模块来进行替换。具体包括 SENet[14]、ECANet[16]和 FcaNet[70]中的压缩模块。为确保实验的公平性，我们在所有对比实验中将 AAP 操作的 s 超参数统一设定为 1，并移除了自注意力和位置注意力的操作。所有被选用的压缩模块旨在生成相同维度的特征描述符，实验所用的基准模型为施加 LADConv(1×) 的 ResNet-18。

表 3-9 LADConv 模型结构中使用不同压缩模块的对比

压缩模块	Top-1 准确率 (%)	算力 (GMac)	参数
Baseline	70.25	1.81	11.69M
SE	73.10	1.84	11.94M
ECA	72.94	1.84	11.69M
Fca	73.07	1.84	11.94M
(Ours)	**73.49**	1.84	12.28M

如表3-9所示，我们设计的压缩模块在与 SE 模块 (实验中第二好的结构) 相当的 FLOPs 和参数配置条件下，实现了 0.39% 的性能提升。我们所选用的对比方法均能够提取通道维度上的特征，这些特征在本质上相当于 1×1 的局部特征。此类方法同样能够在 $k\times k$ 卷积 ($k>1$) 中带来性能提升，因为 1×1 的局部特征与 $k\times k$ 的局部特征之间存在一定的交集。然而，仅依靠 1×1 的局部特征是不足以充分捕捉 $k\times k$ 局部区域内的复杂属性的。通过这一系列实验结果可以看出，引入局部感知特征到卷积核上的注意力机制中，是提升其泛化能力的一种有效且前景广阔的策略。

自注意力及位置注意力操作

表 3-10 LADConv 模型结构中自注意力及位置注意力操作的消融实验

模型	SA	P	$P*$	Top-1 准确率 (%)
Baseline	-	-	-	70.25
+LADConv	-	-	-	73.95
	✓	-	-	74.38
	-	✓	-	74.12
	✓	✓	-	**74.55**
	✓	-	✓	74.15
	-	-	✓	73.84

在压缩模块之后，我们利用 SA(自注意力) 操作来捕捉具有代表性的局部特征之间的相互作用，并通过位置注意力对卷积核捕捉的局部特征进行进一步微调。为了验证这些结构的有效性，我们基于施加了 LADConv(1×) 注意力的 ResNet-18 模型进行实验。当移除 SA 模块时，超参数 s 减小至 1。否则，s 的设置与 ImageNet 实验中的描述保持一致。如表3-10所示，SA 操作和位置注意力操作 (表中定义为 P) 均带来了显著的性能提升。并且，SA 模块的消融实验突显了卷积核注意力中多样化局部特征操作的重要性。此外，为了进一步探究位置注意力与传统空间维度注意力机制的区别，我们进行了额外实验，将 LADConv 中的位置注意力替换为传统的空间注意力 (记为 $P*$)。传统的空间注意力方法首先进行通道维度的全局池化，然后通过 7×7 卷积为每个像素生成注意力系数。然而，从表3-10中的实验结果来看，传统的空间注意力并未能提升模型的准确率。我们认为这一不足主要由两方面原因造成：一是全局池化操作导致了局部特征细节的丧失；二是局部尺寸的不匹配为位置注意力引入了冗余信息，削弱了其有效性。

超参数

表 3-11 LADConv 模型结构中超参数的消融实验

r	s	Top-1 准确率 (%)	算力 (GMac)	参数
8	5	73.52	1.84	12.85M
16	5	73.55	1.84	12.64M
32	5	73.41	1.84	12.45M
16	2	73.35	1.84	12.64M
16	3	73.43	1.84	12.64M
16	8	73.52	1.85	12.64M
16	16	73.58	1.87	12.64M

LADConv 模型在维持基本骨干网络的超参数架构不变的基础上，新增了两个关键的超参数：表示自适应平均池化 (AAP) 后剩余的样本数量 s 和表示降维比例的 r。本实验同样将基于施加了 LADConv(1×) 注意力的 ResNet-18 模型作为出发点。首先，我们对 r 的作用进行了探究。正如表3-11所展现的，$r = 8$ 和 $r = 16$ 的配置产生了相似的结果，并且均优于 $r > 16$ 的配置。因此，在 LADConv 的基础结构中 (应用于 ResNet 系列)，我们默认将 r 设置为 16，在保证引入参数量更少的情况下尽可能提升其性能。超参数 s 反映了局部特征的多样性水平。在

对 ResNet18 各个构建模块 (除了最后一个模块) 进行调整时，我们尝试了不同的 s。根据表3-11展示的实验结果，我们注意到当 s 的值设定在 5 及以上时，相比于 $s < 5$ 的情况，能够实现更加显著的性能提升。然而，较高的 s 同样意味着计算负荷的增加。因此，为了在模型的准确性和计算效率之间达到一个合理的平衡，我们将 s 默认设定为 5。

局部特征尺寸大小

表 3-12 LADConv 模型结构中不同局部特征尺寸大小的影响

局部大小	1	3	5	7
Top-1 Acc (%)	72.88	73.52	73.10	72.95

正如前文所强调的，确保 LADConv 中的局部特征尺寸与卷积核尺寸相匹配对于保证性能至关重要。在表3-12中，我们对展开操作的局部特征尺寸进行了调整，发现一旦局部特征尺寸与核尺寸不一致，模型性能便会显著下降。我们深入探讨了这种不匹配现象的原因：若 LADConv 的局部特征尺寸低于卷积核的尺寸，则会导致模型难以充分理解局部区域的特性；反之，如果局部特征尺寸超出了核尺寸，那么过量的信息就会干扰模型通过压缩模块学习得到的局部特征的准确性。这些发现进一步凸显了调整 LADConv 的局部特征尺寸以匹配原始卷积核尺寸的重要性。

3.3 本章小结

在本章中，我们重新审视了以往的卷积核注意力机制，并观察到了一个显著的局限性。为了解决这一问题，我们提出了 LADConv，这是一种能为卷积引入局部感知注意力的新型方法。我们重新设计了一种压缩结构，以便模型能够有效地捕获卷积核感受野内的局部特征。此外，我们采用自注意力机制来深入挖掘这些局部特征之间的内在联系，旨在保留更多信息丰富的局部特征，以生成更精确的注意力系数。更进一步，我们还在卷积过程中融入了位置注意力机制，对卷积核捕获的不同位置的局部特征进行加权。通过一系列详尽的实验验证，LADConv 展现了其在提升现有卷积架构性能方面的显著优越性，以及具有较大应用价值的潜力。

第四章　细粒度分组特征注意力机制

在上一章中，我们介绍了一种针对卷积核设计的注意力机制 LADConv。继而，本章节将讨论的重心转向特征图上的注意力机制，这是一种更为普遍的注意力研究方向，可以与基于卷积核的注意力机制一并运用于增强网络性能。在结合当前卷积神经网络的发展趋势，并对现有策略的局限进行深入分析后，我们提出了一种新型的细粒度分组特征注意力机制。通过在图像分类和目标检测等视觉任务上的广泛实验验证，证明了该结构能有效提升通用神经网络模型在视觉任务处理中的表现。并以注意力可视化的形式，直观展现了细粒度分组特征注意力机制的优势。

4.1　细粒度分组特征注意力机制

通用卷积神经网络因其强大的特征提取能力成为计算机视觉领域中的一项核心技术。随着对模型性能要求的不断提升，卷积神经网络架构逐渐向更深更复杂的方向发展，尤其是在网络深度增加的同时，产生的特征图也随之在通道数量及语义深度上显著增长。这一变化带来了新的挑战：传统的全局视角注意力机制在处理这些高维、丰富语义的特征图时，往往难以实现精确的特征调整，影响了注意力系数的有效性和模型的最终性能。为应对这一挑战，本文提出了一种细粒度的特征图分组策略，旨在通过对特征图按通道分组并在每个分组内单独计算注意力系数，从而实现对特征的精确调整。此策略的核心优势在于其能够针对每个分组内的特征动态调整权重，更好地捕捉和利用组内的语义关系，同时减少不相关特征的干扰。

针对上述提及的现有基于特征图注意力方法存在的局限，本文提出了一种细粒度的分组特征注意力机制 (Grouped Attention Module with Cross-Level Learning，GAM)。在设计 GAM 时，我们充分考虑了特征图在空间和通道两个维度上的关

键属性，并以 CBAM[15]这种高效融合特征通道和空间维度的注意力机制为基础，对特征图进行细致的分组提取。为了在不显著增加性能开销和模型参数的前提下实施分组注意力计算，我们使用一维卷积替代全连接层来融合组内通道特征。此外，通过引入自注意力机制和残差连接操作，我们对组间的注意力进行全局上下文的增强，从而进一步提升了注意力系数的表征能力。实验结果表明，GAM 在保持参数数量和计算成本相当的情况下，相比现有的特征图注意力机制能够获得更优异的性能表现。

4.1.1 整体模型

GAM 整体模型如图4-1所示。

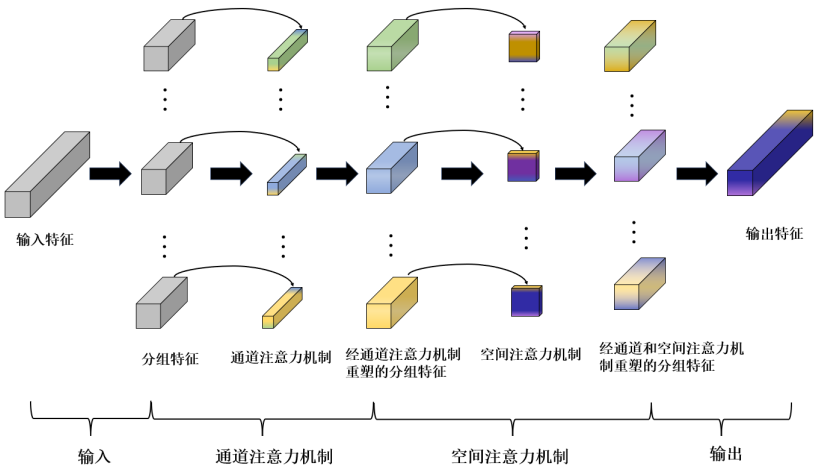


图 4-1 GAM 整体模型示意图

不难看出，GAM 的模型结构可以简单分为以下几个部分：

1. 输入。GAM 接受的输入为卷积神经网络在前向传播过程中产生的中间特征图，这些特征图通常源自网络中一个完整的组件模块的输出。通过 GAM，我们能够对这些输出特征图进行细致的重校准。以 ResNet 网络为例，GAM 的输入特征图便是网络中每一个倒残差结构 (inverted residual block) 计算完成后得到的结果。

2. 通道注意力机制。通道注意力机制致力于对特征通道的重要性进行重新评估。在 GAM 框架中，这一过程涉及将输入特征基于预先设定的阈值分割成若干组，并对每一组内的特征单独计算注意力系数。为了在通道维度上深入挖掘

全局上下文信息，本研究引入了自注意力机制，旨在探索不同组间通道注意力的相互作用与联系。计算完成后，相应的注意力权重将被应用于各个分组特征的通道维度上，以完成通道特征的加权重组。

3. 空间注意力机制。空间注意力机制专注于识别特征空间中的关键区域，以捕获信息丰富的特征。借鉴于 CBAM 注意力模型的架构，我们同样采纳了一种序列化的策略，即先对通道维度进行重塑，再对重塑后的分组特征执行空间注意力的提取。为了尽量减少额外参数的引入，我们对各个分组特征实施了共享的空间注意力提取机制，并通过自注意力机制来强化不同组之间的相互作用。这一过程使得经过空间注意力调整后的特征，能够更加准确地反映出空间中的丰富语义信息区域，从而提升模型对细节的捕捉能力。

4. 输出。GAM 的输出是权值调整完毕后的特征图。在对分组特征施加注意力机制操作之后，我们将这些特征重新组合，恢复至其原始维度，以此作为模块的输出。值得注意的是，在此过程中，引入注意力机制的步骤并不改变各分组特征的维度，因而不会对后续卷积神经网络的前向传播过程产生干扰。通过这种方式，GAM 能够在不影响网络整体架构的前提下，有效地增强特征的表达能力。

接下来会按照数据流动的方向对各个模块的具体实现进行详细的介绍。

4.1.2 特征分组

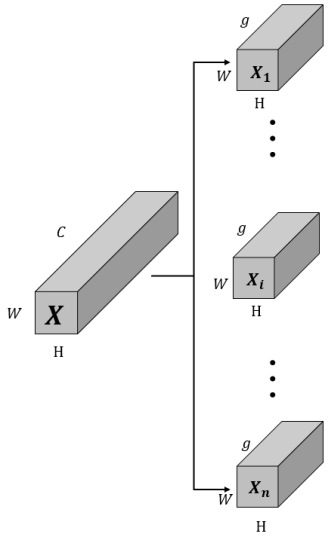


图 4-2 特征分组过程示意图

我们认为，对于富含丰富语义信息的高维通道特征，仅仅通过对整个特征图

执行全局化的操作以获取特征各维度的注意力机制，难以实现对特征权重的精确调整。正如图4-2所示，为了更有效地对这些高维特征实施更精细的注意力机制，我们在提取注意力之前，先沿着通道维度将特征 $X \in \mathbb{R}^{h\times w\times c}$ 按每 g 个通道分组，从而细分为 N 个组别。该过程的形式化表达如下所示：

$$\begin{aligned} N &= \begin{cases} \left\lfloor \frac{c}{g} \right\rfloor & \text{, if } c \geqslant 2g \\ 1 & \text{, otherwise} \end{cases}, \\ X_i &\in \begin{cases} \mathbb{R}^{h\times w\times g} & \text{, if } i < N \\ \mathbb{R}^{h\times w\times (c-g\times (N-1))} & \text{, if } i = N \end{cases}, \end{aligned} \tag{4-1}$$

在处理输入特征时，若通道数 c 小于两倍的分组通道数 $2g$，则不采取分组策略。分组进行时，若发现最后一组特征的通道数不足以形成一个独立的组，则将其与前一个组合并。因此，确定分组总数 N 时需采用向下取整策略。关于超参数 g 的设置将在4.2.4小节内进行讨论。

为了最小化额外参数的增加，进而保留特征注意力机制的即插即用性和轻量级特点，我们设计了一种共享策略，即提取组内通道及空间维度注意力系数时，实现各分组特征共享同一注意力提取网络结构。这一策略的具体实现方法是将分组特征的组数维度融合进批次 (batch) 维度，此操作不仅简化了模型架构，还实现了注意力机制相关计算的并行处理，有效降低了模型训练和推理阶段的时间开销。

4.1.3 通道注意力机制

GAM 的通道注意力机制模块如下图所示：

我们将特征按通道进行分组，并针对每个分组内部执行通道注意力提取操作，有效地加深了模型对特征通道间关系的理解。考虑到每个通道在特征图中扮演的角色相当于一个特征检测器[71]，分组后的通道注意力操作使得模型能够在更细粒度层面上关注输入图像中有意义的信息。不同组内的通道可能捕获到相互补充或独立的特征信息，与直接在高维特征全局层面上学习通道间交互相比，这种对每个分组独立融合通道信息的方法能够更为有效地凸显出各分组中重要的特征检测器，从而提升模型对输入图像的理解能力。

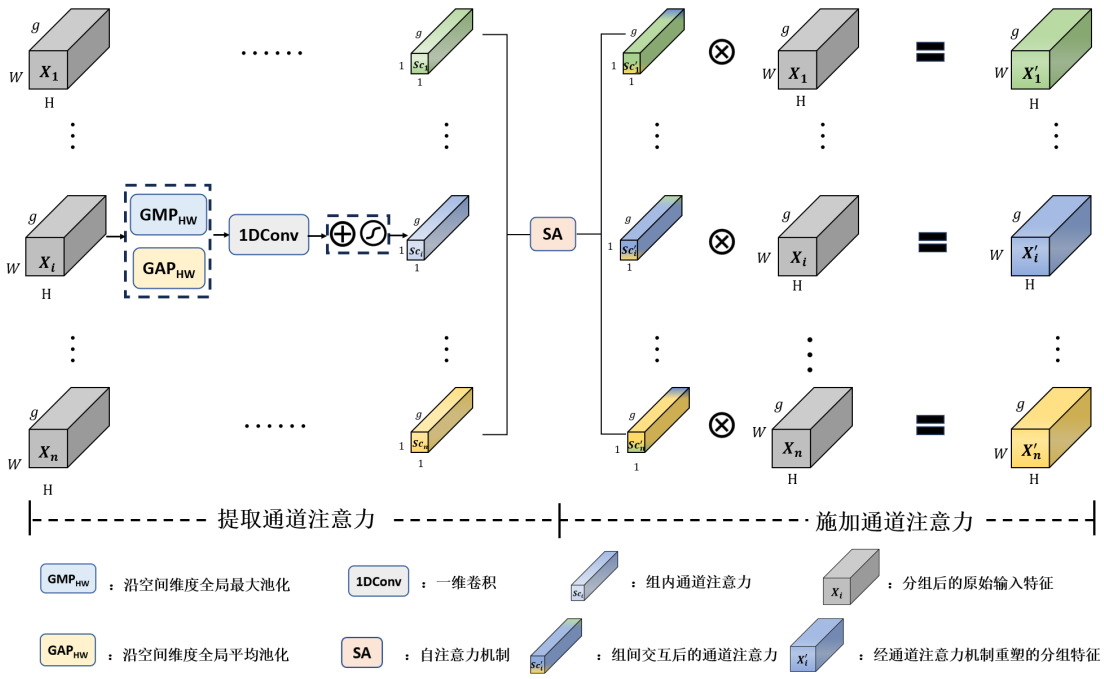


图 4-3 通道注意力模块结构示意图

具体而言，在处理各分组特征时，我们首先对特征图的空间维度执行平均池化和最大池化操作，以此聚合空间信息，从而生成两种不同的空间上下文描述符 F_{avg}^{c} 和 F_{max}^{c}，为了尽量减少分组操作可能引入的模型复杂度，并更有效地确立权重向量与输入通道之间的映射关系，我们采用一维卷积来替换传统采用的两个全连接层的降维激励模块。这样的处理不仅简化了模型结构，还增强了模型的计算效率。因此，组内通道注意力的提取过程可以通过以下公式形式化描述：

$$
\begin{aligned}
F_{\text{avg}}^{c} &= \text{GAP}^{HW}(X), \\
F_{\text{max}}^{c} &= \text{GMP}^{HW}(X), \\
s_c &= \sigma\left(\text{Conv1D}\left(F_{\text{avg}}^{c}\right) + \text{Conv1D}\left(F_{\text{max}}^{c}\right)\right),
\end{aligned}
\tag{4-2}
$$

为了使公式简洁，上述计算过程省略了组号下标 i。其中，GMP^{HW} 和 GAP^{HW} 分别代表空间维度的最大池化和平均池化操作。Conv1D 是我们采用的形状为 k 的一维卷积核，以模拟分组内部的跨通道相互作用。这里，参数 k 起到了决定交互范围的关键作用，并会根据每个分组的通道维数 g 进行自适应调整，以确保

交互效果与分组结构相匹配：

$$k = \psi(g) = \left| \frac{\log_2(g)}{\gamma} + \frac{b}{\gamma} \right|_{\text{odd}}, \tag{4-3}$$

超参数 γ 和 b 参考 ECANet[16]中的取值分别设置为 2 和 1。$|t|_{odd}$ 表示最接近 t 的奇数值。

经过上述过程，我们得到了 n 个分组特征的组内通道注意力机制 s_c。为了强化各分组特征内部通道注意力之间的相互作用，以及增强通道维度上特征间的长距离依赖性，本研究采用了一种计算各分组内通道注意力自注意力系数的方法，进而对通道注意力权重进行了精细化的调整。此外，为了最大程度地保留每个分组内通道注意力机制的局部特性，我们采纳了残差连接的策略，将得到的自注意力系数直接叠加到原始的通道注意力系数之上。这种方法保证了在增强分组特征通道注意力的全局信息同时，不丢失每个通道注意力独有的局部信息。将各组内通道注意力机制重组为张量 $S_c = \left[s_c^1, s_c^2, \cdots, s_c^n\right]$，$S_c \in \mathbb{R}^{n \times g}$，自注意力机制计算如下式所示：

$$\begin{aligned} S_c' &:= \text{LN}(S_c), \\ \text{SA}(S_c') &= \text{Softmax}(\frac{W_Q\left(S_c'\right) W_K\left(S_c'\right)^\top}{\sqrt{n}}) W_V\left(S_c'\right), \\ S_c &:= \text{SA}(S_c') + S_c, \end{aligned} \tag{4-4}$$

其中，LN 表示层归一化操作 (Layer Normalization，LN)，SA$(\cdot)$ 代表关于 $\cdot$ 的自注意力系数。$W_Q(\cdot)$，$W_K(\cdot)$ 和 $W_V(\cdot)$ 对应于输出尺寸为 n 的全连接层，用于完成对 S_c' 的线性变换操作。最终，我们将计算得出的新的组内通道注意力系数应用于相应的分组特征通道维度之上，以此来优化和调整各分组特征内部的通道关系：

$$X' = s_c X, \tag{4-5}$$

同样地，上式也省略了序号标识 i。

从更直观的层面来讲，在 GAM 通道注意力机制的计算过程中，采用特征分组的设计主要旨在对通道注意力机制网络结构设计中常见的压缩-激励模块中的激励部分进行优化。因为压缩模块旨在压缩特征空间维度的信息，即 $H \times W$ 的

空间池化尺度不受分组操作影响，而在之后的激励操作中，我们将全局的高维通道维度信息细化为多个局部低维通道信息的集合，实现对通道注意力机制的更细致提取。

4.1.4 空间注意力机制

GAM 的空间注意力机制模块如下图所示：

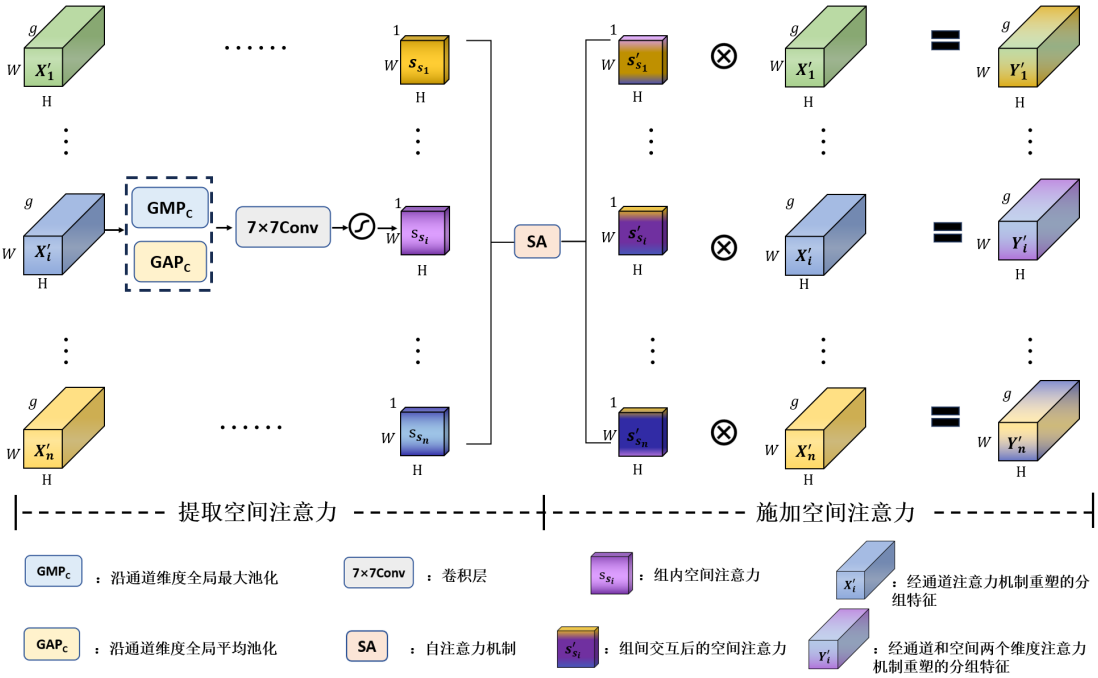


图 4-4 空间注意力模块结构示意图

相较于通道注意力机制关注给定输入特征中的关键信息是什么，空间注意力机制专注于识别关键信息所在的位置，起到与通道注意力互补的作用。沿通道维度进行池化操作被验证为一种有效的手段，能够突出显示信息的重要区域[72]。而不同于传统方法在整个通道维度进行全局池化，GAM 中的空间注意力机制的池化范围受限于各分组特征的通道数 g，这种有针对性的局部池化策略不仅加强了保留信息与特征分组间的紧密联系，还显著提高了空间注意力机制在捕捉细节上的精确性与效率。

我们对经通道注意力机制调整后的各分组特征应用平均池化和最大池化操作，并将产生的结果连接起来以生成有效的特征描述符。对于连接后的特征描述符，我们参照 CBAM 中的操作，运用一个尺寸较大的卷积层来生成各分组特征的空间注意力系数 $s_s \in \mathbb{R}^{H \times W}$ 以此编码出哪些位置应被强调或抑制。具体而

言，我们采用两种池化操作来整合特征图中的通道信息，从而产生两个二维的描述符：$F^{\mathbf{s}}_{\text{avg}} \in \mathbb{R}^{1\times H\times W}$ 和 $F^{\mathbf{s}}_{\text{max}} \in \mathbb{R}^{1\times H\times W}$ 分别代表对各个分组特征通道维度执行的平均池化和最大池化所得到的描述符。接着，我们将这两个描述符通过一个 7×7 的标准卷积层进行合并和卷积处理，以此生成我们的二维空间注意力图。形式化表达如下：

$$\begin{aligned} F^{s}_{\text{avg}} &= \text{GAP}^{c}(X'), \\ F^{s}_{\text{max}} &= \text{GMP}^{c}(X'), \\ s_s &= \sigma\left(\text{Conv}_{7\times 7}\left(\left[F^{s}_{\text{avg}}; F^{s}_{\text{max}}\right]\right)\right). \end{aligned} \tag{4-6}$$

经过上述过程，我们得到了 n 个分组特征的组内空间注意力机制 s_s，同样为了使其简洁，我们在公式4-6中省略了组号下标 i。为了进一步增强各分组特征内部空间注意力的相互作用，并在最大限度上保留每个分组内部空间注意力机制的局部特性，我们采用了与第4.1.3节中相似的自注意力机制及残差连接的策略来对分组特征的空间注意力系数进行细致调整。首先将各组内空间注意力机制重组为张量 $S_s = [s^1_s, s^2_s, \cdots, s^n_s]$，$S_s \in \mathbb{R}^{n\times H\times W}$，自注意力机制计算如下：

$$\begin{aligned} S'_s &:= \text{LN}\left(S_s\right), \\ \text{SA}\left(S'_s\right) &= \text{Softmax}\left(\frac{W_Q\left(S'_s\right) W_K\left(S'_s\right)^{\top}}{\sqrt{n}}\right) W_V\left(S'_s\right), \\ S_s &:= \text{SA}\left(S'_s\right) + S_s, \end{aligned} \tag{4-7}$$

公式中各符号表意与公式4-4保持一致，这里不再过多赘述。最终，我们将计算得出的新的组内空间注意力系数应用于相应的分组特征空间维度之上，以此来优化和调整各分组特征内部的空间位置关系：

$$Y' = s_s X'. \tag{4-8}$$

4.1.5 特征重组

如图4-5所示，在完成对各组特征通道及空间维度的注意力机制施加后，将这些特征按照原分组顺序重新组合，恢复至其原始特征维度。由于引入注意力

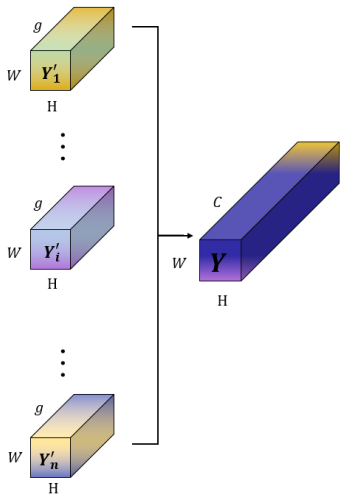


图 4-5 特征重组过程示意图

机制的网络结构中不涉及对分组特征自身维度的变换，因而恢复至其原始特征维度不需要引入维度转换结构，并且不会对后续卷积神经网络的前向传播流程产生干扰。这一过程确保了特征在经过细粒度注意力机制的调整后，能够无缝地融入后续的网络结构中，同时维持了数据的完整性和整个网络结构的连贯性。

4.2 实验与分析

为了全面评估分组注意力机制 GAM 在实际视觉任务中的性能，本研究沿用了类似于3.2节中的实验设置，选取了 ImageNet、CIFAR-100 以及 MS COCO 这三个广泛应用于视觉任务研究的标准数据集进行实验验证。这些数据集的详细介绍已在3.2.1小节中提供，因此本节不再赘述其具体内容。同时，对于图像分类和目标检测任务的实验评价指标，本文已在第3.2.2节中提供了详尽的形式化描述和解释。

4.2.1 ImageNet 对比实验

实验设置

在 ImageNet 数据集上的实验中，本文将分组注意力机制 (GAM) 集成至多种主流的通用卷积神经网络架构进行评估，包括 ResNet-18、ResNet-34、ResNet-50、ResNet-101[8]、ResNeXt-50[67]以及 MobileNetV2[42]等。同时，本研究还对几

种广泛采用的特征图注意力机制进行了深入的比较分析，这些注意力机制包括SENet[14]、CBAM[15]、ECA[16]和 SRM[73]注意力。为了确保实验结果的公平性和可比性，我们在 PyTorch 框架下对所有待比较的方法采取了统一的实验设置，并且对比方法均通过使用各自作者提供的公开代码进行了精确复现。这一标准化的实验流程旨在提供一个公正的基准，以准确评估 GAM 在提升网络性能方面的效果。

在本文中，对于 ImageNet 数据集上基于特征图的注意力机制模型的实验设置与3.2.3节中基于卷积核的注意力机制在该数据集上的对比实验设置几乎保持一致。然而，为了保证比较结果的公正性，我们在特征图注意力机制模型的训练过程中做了一些调整：首先，我们没有采用温度退火策略，该策略主要用于优化聚合多个静态卷积核的注意力系数中的 Softmax 激活函数，鉴于特征图注意力机制中不存在这一结构，故排除了此策略的使用。同时，我们也去除了模型中的 dropout 策略。GAM 中独特的分组特征组内通道数 g 在 ImageNet 数据集的实验中统一设置为 32，以确保在进行模型计算时，不需要对分组特征通道数做特殊处理。

实验结果

GAM 在 ImageNet 验证集上的对比实验结果如表格4-1所示。除了在 ResNet-18 架构上的 Top-5 准确率略低于 SE 结构外 (下降了 0.04%)，GAM 在所有其他模型架构上均实现了最优的 Top-1/Top-5 准确率表现。具体而言，与次优的特征注意力结构相比，GAM 在 ResNet-18 上的 Top-1 准确率提高了 0.07%，在 ResNet-34 上的 Top-1/Top-5 准确率分别提高了 0.09%/0.06%，在 ResNet-50 上的 Top-1/Top-5 准确率提升了 0.49%/0.11%，在 ResNet-101 上的 Top-1/Top-5 准确率分别提高了 0.83%/0.34%，在 ResNeXt-50 上的 Top-1/Top-5 准确率分别提升了 0.29%/0.17%，而在 MobileNetV2 上的 Top-1/Top-5 准确率分别提高了 0.07%/0.04%。通过上述实验表现可以证明，GAM 结构能够在不显著增加计算负担和参数数量的情况下，有效地提升特征注意力机制在通用卷积神经网络中的性能表现。

在这些结果中尤其值得注意的是，GAM 对比次优的特征注意力结构在 ResNet-50 和 ResNet-101 上的性能提升相较于 ResNet-18 和 ResNet-34 更为显著。这一现象揭示了一个重要的观点：对于层次结构较深的通用卷积神经网络而言，GAM

表 4-1 GAM 在 ImageNet 验证集上基于通用卷积神经网络系列模型的测试结果

模型	Top-1 准确率 (%)	Top-5 准确率 (%)	算力 (GMac)	参数 (MB)
ResNet-18	70.33	89.58	1.82	11.69
+SE	71.19	**90.21**	1.82	11.78
+CBAM	71.24	90.04	1.82	11.78
+ECA	70.71	89.85	1.82	11.69
+SRM	71.09	89.98	1.82	11.69
+GAM(ours)	**71.31**	90.17	1.82	11.80
ResNet-34	73.75	91.60	3.67	21.80
+SE	74.32	91.99	3.67	21.95
+CBAM	74.41	91.85	3.67	21.96
+ECA	74.03	91.73	3.67	21.80
+SRM	74.49	92.01	3.67	21.81
+GAM(ours)	**74.58**	**92.05**	3.67	22.01
ResNet-50	76.34	93.12	4.11	25.56
+SE	77.51	93.74	4.12	28.07
+CBAM	77.63	93.88	4.12	28.09
+ECA	77.17	93.52	4.12	25.56
+SRM	77.51	93.06	4.11	25.59
+GAM(ours)	**78.12**	**93.99**	4.13	29.33
ResNet-101	77.82	93.85	7.83	44.55
+SE	78.39	94.13	7.85	49.29
+CBAM	78.57	94.18	7.85	49.33
+ECA	78.46	94.12	7.84	44.55
+SRM	78.58	94.15	7.83	44.68
+GAM(ours)	**79.41**	**94.29**	7.88	51.98
ResNeXt-50(32x4d)	77.47	93.52	4.26	25.03
+SE	77.96	93.93	4.27	27.54
+CBAM	78.06	94.07	4.27	27.56
+ECA	77.74	93.87	4.27	25.03
+SRM	78.04	93.91	4.26	25.06
+GAM(ours)	**78.35**	**94.24**	4.28	29.02
MobileNetV2	71.90	90.51	0.31	3.50
+SE	72.46	90.85	0.31	3.53
+CBAM	72.49	90.78	0.32	3.54
+ECA	72.01	90.46	0.31	3.50
+SRM	72.32	90.70	0.31	3.51
+GAM(ours)	**72.56**	**90.89**	0.32	3.56

能够更加有效地捕捉特征组间信息的细微差距，并据此优化调整特征的注意力分配。原因在于，在较深的网络结构中，生成的特征图不仅通道数更多，而且各个特征组之间的信息差异性更加突出。GAM 通过采用更为精细化的注意力机制处理这些特征，能够更加精确地识别和利用这些差异性，进而显著提升模型在处理复杂视觉任务时的性能。这一策略不仅增强了模型对细节信息的捕捉能力，也为深层卷积神经网络的性能优化提供了有效途径。

4.2.2 CIFAR-100 对比实验

实验设置

为了进一步探究 GAM 在图像分类任务上的性能，我们在 CIFAR-100 数据集上展开了补充实验。本轮实验选取了包括 ResNet-20、ResNet-46、ResNet-110 以及 MobileNetV2(0.75×) 在内的一系列更适合 CIFAR-100 数据集的轻量级网络模型。优化策略采用了动量为 0.9 的 SGD 优化器，设置批量大小为 128，权重衰减系数为 0.0 005。所有网络均在单个 GPU 上进行训练，学习率从 0.1 开始，并在第 32 000 次和第 48 000 次迭代时分别降低至原来的十分之一，直至第 64 000 次迭代停止调整。鉴于这些模型的网络结构相对简单，我们将分组特征的通道数从 32 调整为 8，这一调整旨在更清晰地展示特征分组及其注意力机制施加的效果。同时，考虑到在 CIFAR-100 数据集上，卷积神经网络的性能显著受到随机初始化的影响，我们采用固定种子进行五次重复实验，并取平均值作为最终结果，以保证实验的稳定性和结果的可靠性。

实验结果

表 4-2 GAM 在 CIFAR-100 验证集上基于通用卷积神经网络系列模型的测试结果

模型	Top-1 准确率 (%)	模型	Top-1 准确率 (%)
ResNet-20	68.88	ResNet-56	72.24
+SE	69.45	+SE	72.84
+CBAM	69.47	+CBAM	72.47
+ECA	68.89	+ECA	72.45
+GAM(ours)	**69.52**	+GAM(ours)	**72.96**
ResNet-110	75.54	MobileNetV2(0.75×)	67.32
+SE	76.56	+SE	67.54
+CBAM	76.54	+CBAM	67.79
+ECA	76.33	+ECA	67.24
+GAM(ours)	**76.81**	+GAM(ours)	**67.92**

GAM 在 CIFAR-100 验证集上的性能比较结果如表格4-2所示。不难看出，GAM 在四种轻量级的通用卷积神经网络架构中均取得了最佳的 Top-1 准确率提升。具体来看，GAM 分别在 ResNet-20、ResNet-56、ResNet-110 以及 MobileNetV2(0.75×) 上较次优的特征注意力结构相比分别实现了 0.05%、0.12%、0.25% 和 0.13% 的准确率提升。CIFAR-100 上的实验结果同样揭示了一个规律：拥有更深层次网络

结构的模型从 GAM 的应用过程中获得了更显著的性能提升。在 CIFAR-100 数据集上的实验再次印证了 GAM 通过分组施加注意力机制的有效性。

4.2.3 MS COCO 对比实验

实验设置

为了探究 GAM 在目标检测视觉任务上的表现，我们选择了较为流行的 Faster R-CNN[64]和 Mask R-CNN[74]目标检测框架，并结合特征金字塔网络作为连接部件进行了一系列实验。实验所选用的骨干网络是在 ImageNet 数据集上图像分类视觉任务中表现优异的 ResNet-50 和 ResNet-101 网络，用于比较的注意力机制是模型结构较为相似的 CBAM[15]特征注意力。这些网络模型均先在 ImageNet1K 数据集上进行预训练，随后通过微调 (fine-tuning) 过程迁移到 COCO 数据集上。为了保证实验比较的公平性，所有模型的训练和评估均借助 MMDetection 框架开展，根据线性缩放规则[75]，起始学习率设置为 0.01。其余超参数设置遵循各检测器的默认配置。

实验结果

表 4-3 GAM 在 COCO 验证集上基于 Faster RCNN 和 Mask RCNN 模型的测试结果

Framework	BackBone	mAP	Framework	BackBone	mAP
Faster RCNN	ResNet-50	37.8	Mask RCNN	ResNet-50	38.1
	+CBAM	39.3		+CBAM	39.9
	+GAM(ours)	**39.4**		+GAM(ours)	**39.9**
	ResNet-101	39.8		ResNet-101	40.3
	+CBAM	41.2		+CBAM	41.6
	+GAM(ours)	**41.4**		+GAM(ours)	**41.8**

正如表格4-3所示，将 CBAM 或我们提出的 GAM 特征注意力机制整合进 ResNet 模型中，能显著增强目标检测视觉任务相关基准模型的性能。具体而言，GAM 特征注意力机制为上述各类目标检测框架带来的性能提升甚至略微超过 CBAM。通过在 COCO 数据集上的目标检测视觉任务相关实验表明，我们所提出的 GAM 分组特征注意力机制带来的优化并不受限于特定的数据集或特殊的卷积神经网络结构，而是能够客观地提升通用卷积神经网络在视觉任务中的表

现。这一发现证实了 GAM 作为一种符合视觉注意力机制定义的即插即用网络结构，在视觉识别领域中具备实用性和有效性。

4.2.4 消融实验

通过第4.1节对 GAM 网络结构的详细阐述，我们可以将 GAM 架构解构为四个关键组成部分：分组操作、特征通道维度的注意力机制、特征空间维度的注意力机制以及自注意力机制，并在分组操作中引入了分组特征通道数 g 这一重要的超参数。由于不实施分组操作本质上等同于将分组特征通道数设为与输入特征通道数相同，因此本节专门就通道注意力、空间注意力、自注意力机制以及超参数 g 展开了两项消融实验。这些实验旨在深入探讨各组成部分对模型性能的具体贡献。并随后通过注意力可视化技术，直观地展示了 GAM 与其他注意力机制之间的差异，以便于更好地理解 GAM 在提升模型识别能力方面的独特优势和作用机制。

GAM 各模块消融

为了深入研究 GAM 中不同模块对提升通用卷积神经网络性能的具体贡献，我们以 ResNet-50 为基准模型开展了关于通道注意力机制、空间注意力机制和自注意力机制模块的消融实验。数据集选用 ImageNet1K，相关实验设置与4.2.1节中保持一致。

表 4-4 GAM 模型结构中各模块的消融实验

模型	Channel	Sptial	SA	Top-1 准确率 (%)
ResNet-50	-	-	-	76.34
+GAM	✓	-	-	77.29
	-	✓	-	77.14
	✓	✓	-	77.42
	✓	-	✓	77.64
		✓	✓	77.57
	✓	✓	✓	**78.12**

消融实验结果如表4-4所示，我们可以清晰地观察到各个模块之间显示出了较强的互补能力。此外，与分组空间注意力机制相比，分组通道注意力机制对于提升通用卷积神经网络的性能贡献更为显著。而自注意力机制模块则在这两者的基础上进一步增强了性能表现，此结果进一步证实，除了促使注意力机制专

注于分组特征的内部特点外，强化组间联系，以促进不同分组间的信息补充，对于设计更精细化的注意力机制同样至关重要。

分组特征通道数

表 4-5 GAM 模型结构中关于分组特征通道数的消融实验

g	Top-1 准确率 (%)	算力 (GMac)	参数 (MB)
c	77.56	4.12	27.49
64	77.94	4.12	28.75
32	78.12	4.13	29.33
16	**78.13**	4.14	30.06
8	77.79	4.15	30.71

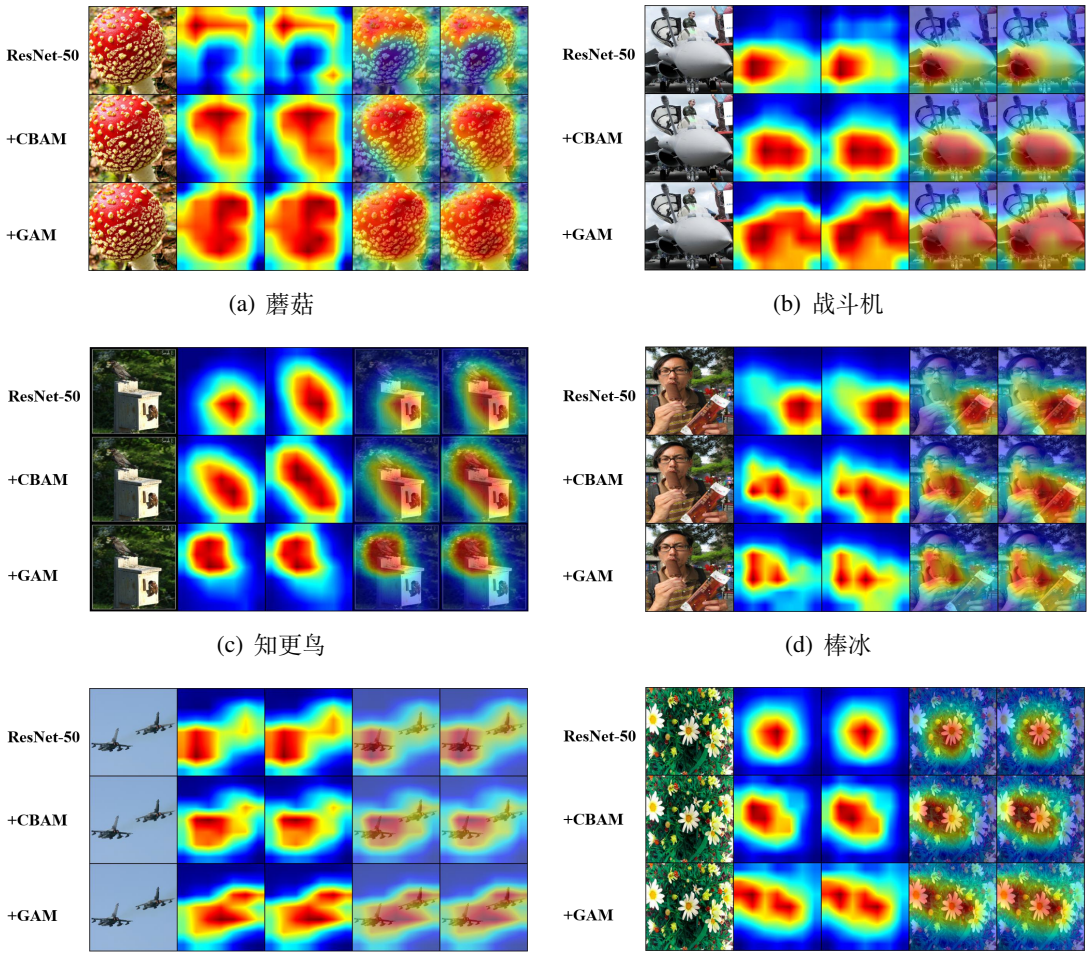
在 GAM 中，分组特征的通道数 g 直接决定了输入特征的分组数量，进而影响到通用卷积神经网络模型的计算复杂度和参数规模。以 ResNet-50 作为参考基准，我们在 ImageNet1K 数据集上探讨了不同 g 值对卷积神经网络模型的影响，结果汇总在表4-5中。表中，c 代表分组特征通道数 g 与输入特征通道数 c 相等，即未对输入特征进行分组直接应用 GAM 注意力机制的情况。实验结果表明，当 g 设为 32 和 16 时，GAM 能够为通用神经网络带来最显著的性能提升。相较而言，当 g 取值为 16 时，虽然性能有所提升，但与之引入的额外计算量和参数不相匹配，因此在我们的标准实验配置中选用了 g 为 32 的情况。对于 g 取值为 8 时，性能提升不甚明显，推测原因可能是较小的通道数限制了能够携带的信息量，从而影响了注意力机制的有效提取。

注意力机制可视化

为了通过实际的例子来揭示注意力机制如何使通用卷积神经网络更加专注于其关注的区域，我们采用了 Grad-CAM(Gradient-weighted Class Activation Mapping)[76]和 Grad-CAM++[77]技术，以便更加直观地理解卷积神经网络是如何根据特定输入图像进行分类决策的。这两种技术提供了一种可视化的方法，通过它们，我们可以清晰地观察到网络在做出分类判断时，哪些区域被视为决策的关键，进而更好地理解模型的判断依据和注意力机制的作用效果。

Grad-CAM 是一种类激活映射方法，它利用卷积神经网络 (本次实验中选用 ResNet-50 作为基准模型) 最后一个卷积层的梯度信息来理解网络对于给定图像

进行特定分类决策的依据。具体而言，Grad-CAM 通过计算目标类别相对于特征图的梯度，并将这些梯度加权的特征图线性组合，生成一个粗略的热力图。这个热力图能够突出显示对于模型进行特定类别预测至关重要的区域，为模型的决策过程提供了一种直观的视觉解释。而 Grad-CAM++ 则在 Grad-CAM 的基础上进一步引入了二阶和三阶梯度计算，以增强梯度信息的丰富度和准确性。如图4-6所示，实验中选取的每张图片 (ImageNet1K 验证集中随机挑选) 及其相应的热力图展示了卷积神经网络模型在识别过程中最为关注的区域。通过比较 Grad-CAM 和 Grad-CAM++ 生成的热力图 (分别位于第二列和第三列)，以及将这些热力图与原始图像叠加后得到的结果 (分别位于第四列和第五列)，我们可以更加清晰地识别出通用卷积神经网络关注焦点所在的区域，从而更深入地理解网络在图像分类任务中的行为模式和决策依据。



(a) 蘑菇 (b) 战斗机

(c) 知更鸟 (d) 棒冰

(e) 飞机 (f) 雏菊

图 4-6 注意力机制可视化对比示意图

通过上述实验观察，我们发现 GAM 分组特征注意力机制展现出以下几个显

著特征：

1、全面捕捉物体特征：对比4-6(a) 和4-6(b) 两幅子图可见，未加入注意力机制的卷积神经网络往往只能捕捉到物体的部分区域。而引入 CBAM 之后，模型的关注范围虽有所扩展，但依然遗漏了一些关键细节。而 GAM 的引入，使得网络模型能够在更全面的程度上捕捉到物体的整体特征，有效降低了误分类的可能性。

2、排除无关背景的干扰：从4-6(c) 和4-6(d) 两幅子图可以清晰地看到，未引入注意力机制的网络受到背景中无关物体的明显干扰 (如 c 图中的鸟巢和 d 图中的包装纸)。尽管 CBAM 能够一定程度上使网络聚焦于正确的物体，但仍然未能完全排除无关背景干扰。GAM 则有效地定位了目标物体区域，使得模型能够准确判断。

3、同时关注多个目标物体：从4-6(e) 和4-6(f) 两幅子图可见，初始模型倾向于集中注意力于图片中的单一物体上。CBAM 的加入虽然扩大了关注范围，但并未能全面覆盖第二个物体。GAM 则实现了对多个物体的均衡关注，显著提升了模型在处理含有多个目标物体的视觉任务中的表现力。

综上所述，GAM 通过在分组层面实施注意力机制，不仅增强了通用卷积神经网络捕捉目标物体细节的能力，而且有效减少了非目标区域干扰，显著提升了模型在复杂视觉任务中的性能表现，彰显了其在精细化视觉识别任务中的应用价值。

4.2.5 卷积核注意力机制和特征图注意力机制

回顾第三章中提出的具有局部感知能力的卷积核注意力机制与本章提出的分组特征图注意力机制，我们注意到，虽然直接对卷积核施加注意力机制在性能提升方面较为显著，但基于特征的注意力机制引入的参数数量和计算量相对更少，且实施该机制的操作过程更加直接和简便，无需为通用卷积神经网络中各种特殊卷积核量身定做注意力实施策略。为探究这两种注意力机制是否具有互补性，我们选取 ResNet-50 作为基准模型，在 ImageNet1K 数据集上进行了一系列补充实验，其中具体的实验配置与第4.2.1节中的设置相同。

如表4-6所示，我们提出的 GAM 特征注意力机制能够在 LADConv 卷积核注意力机制的基础上，进一步提升通用卷积神经网络在视觉任务上的性能。特征注

表 4-6 GAM 和 LADConv 在 ImageNet1K 验证集上基于 ResNet-50 的测试结果

模型	Top-1 准确率 (%)	Top-5 准确率 (%)	算力 (GMac)	参数 (MB)
ResNet-50	76.34	93.12	4.11	25.56
+GAM	78.12	93.99	4.13	29.33
+LADConv(4x)	79.91	94.13	4.32	91.41
+GAM&LADConv(4x)	**80.14**	**94.20**	4.34	95.18

意力机制通过重新分配特征权重，引导网络更加专注于图像中的关键区域；同时，卷积核注意力机制则使得模型能够根据不同的输入动态地调整其网络参数，从而更有效地捕捉和利用特征信息。这两种机制的结合使用，展现了注意力机制对于增强通用卷积神经网络视觉任务表现所具有的巨大潜力。

4.3 本章小结

在本章中，我们以基于特征的注意力机制为出发点，探讨了当前方法的潜在改进空间，并提出了一种细粒度的分组特征注意力机制 GAM。GAM 通过特征分组的策略，分别对各分组特征组内的通道和空间维度施加注意力，进而利用自注意力机制强化注意力系数的组间联系，为具有高维通道的特征实现更精细化的注意力处理。随后通过在标准的视觉任务数据集上的测试结果表明，GAM 能显著提升通用卷积神经网络的性能，并通过消融实验和注意力可视化技术直观证明了 GAM 结构的有效性。最后，我们结合上一章介绍的 LADConv 卷积核注意力机制，进一步探索了两者之间的互补性，从而全面验证了注意力机制在提升卷积神经网络处理视觉任务性能方面的关键作用。

第五章　视觉注意力机制在垃圾检测系统中的应用

在第三章和第四章中，我们分别探讨了基于卷积核的注意力机制 LADConv 与基于特征图的注意力机制 GAM。本章节，我们基于 Yolov8s 目标检测框架设计了一套完整的垃圾检测系统。在该系统中，我们将 LADConv 和 GAM 这两种注意力机制的不同组合方案融入原有的骨干网络结构中，以在资源受限的边缘设备上发挥出模型最好的性能。值得强调的是，我们设计的系统并不局限于特定模型；相反，它允许根据实际任务需求灵活替换注意力模型，并根据数据集进行训练与测试。

5.1　系统研发背景

路面垃圾检测作为城市环境保护的一项关键措施，在维护城市清洁和提升居民生活质量方面发挥着至关重要的作用。在传统的环卫工作模式中，垃圾清理工作的开展极大程度上依赖于工人们手持各种清洁工具或驾驶环卫车辆进行实地巡查。这种依靠人力进行路面清洁的方式不但效率相对较低，还很难对城市中宽阔和复杂的环境做到全面覆盖。随着科技的发展和智能化技术的应用，如今有望通过自动化的路面垃圾检测系统来优化这一传统模式。通过将深度学习技术部署在智能监控设备上，能够实现对城市路面垃圾的自动检测和分类，不仅可以提升环卫工作的效率，还能显著减少对人力资源的依赖，推动城市环境保护工作向更加智能化、高效化的方向发展。

目前，市场上常见的视频监控系统大多基于 Yolo 等静态目标检测技术，这类系统一旦部署后便很少进行更新。然而，路面垃圾检测是一项长期而持续的任务，在短期内难以获得特定街道的大量垃圾数据，系统需要支持即使是非人

废纸盒

饮料瓶

图 5-1 两种常见的路面垃圾

工智能专业领域的用户也能够持续搜集数据并更新模型，以应对复杂多变的路面环境和垃圾类别。并且，考虑到路面垃圾检测系统需要在各种性能各异的边缘智能设备上运行，这些设备对于模型大小和处理速度的要求有着明显的差异。因此，根据具体的硬件条件选择最适宜的注意力机制，以最大化现有目标检测网络的性能变得尤为重要。

为了克服现有系统的局限性，我们设计了一个涵盖数据标注、自动化训练测试到最终边缘设备部署全流程的目标检测系统。考虑到不同的硬件环境需求，我们设计了一系列注意力机制配置方案，包括 GAM、LADConv(1×)、LADConv(4×) 以及 GAM+LADConv(4×) 等，并且用户能够在使用过程中持续更新和优化模型，保障系统能够长期有效地满足城市路面垃圾检测的需求。

5.2 系统设计

在本节中，我们对垃圾检测系统进行了全面的需求分析，深入探讨了各项需求所面临的挑战及期望达成的目标效果。基于这一分析，我们进一步设计了一个完善的系统架构，旨在全面满足垃圾检测任务的各项需求。

5.2.1 系统需求分析

需求分析是系统开发的基石，精确捕捉需求能帮助开发人员避免在错误的方向上浪费大量时间与资源。针对路面垃圾检测系统，我们基于用户习惯、主观反馈及实际应用场景，细致梳理了系统的关键需求如下：

1. 标注工具集成：针对用户通过互联网下载或现场拍摄获取到的图像及视

频资料，这些素材多数情况下未包含标注信息。而现有的数据标注工具存在一定的操作门槛，尤其对于没有技术背景的用户来说，使用起来可能会遇到挑战。因此，我们设计了一种简化的数据标注流程。该流程旨在降低标注过程的难度，确保用户能够轻松、准确地对图像中的垃圾种类及其位置进行有效记录。

2. 智能数据解析：针对系统识别得到的垃圾数据，我们应用了一套高效率的自动化数据分析算法。该算法旨在对垃圾的种类、数量、常见出现地点以及出现时间等关键指标进行精细化统计。通过深入分析这些关键数据，环卫管理部门能够更科学地进行人力资源的分配与调度安排，从而在根本上提升城市清洁工作的整体效率和效果。

3. 持续学习机制：考虑到当前街道的清洁状况普遍良好，使得路面垃圾的样本相对稀少，导致最初构建的数据集规模较小且样本分布不均。为此，我们设计了一个持续学习的框架，该框架能够持续接纳新采集的样本数据，并利用这些数据对模型进行迭代更新。通过这种方式，系统能够逐步提升模型的检测精度，确保模型检测能力的持续精进。

4. 模型仓库管理：随着模型不断迭代更新，系统将积累大量训练结果，以适配多样化的应用场景。针对这一问题，我们开发了一套用户友好的模型管理工具。该工具允许用户轻松执行模型版本的筛选、部署、保留或移除操作，从而确保了系统资源得到高效且合理的分配。

5. 边缘计算兼容性：鉴于边缘计算设备的多样化特点及其带来的计算资源限制不同，我们采取了多种注意力机制的组合优化方案，以确保模型能够适应各种不同规模的部署需求。为了高效地协调边缘设备与中心服务器之间的数据交流，我们设计并实施了一套先进的数据通信机制。该机制确保了模型参数与检测结果能够在实时环境下进行同步，大幅提升了整个系统的反应敏捷性与数据处理效率。

通过上述的需求分析，我们旨在打造一个高效、自动化、易于操作的路面垃圾检测系统，以实现在实际城市环境中的广泛应用。

5.2.2 系统架构设计

在确定了系统所需满足的核心需求后，我们深入分析并筛选了潜在的技术方案，构筑出一套理想的系统总体设计。正如图5-2所示，该垃圾分类系统主要

由边缘计算层与数据处理层两大核心部分构成。边缘计算层主要任务是运行经过精细调优的注意力增强目标检测算法，确保对各类垃圾进行高效精准的识别；数据处理层则承担着与终端用户进行有效交互的职责。接下来，我们将逐一探讨系统内各个关键组件的功能定位及其重要性：

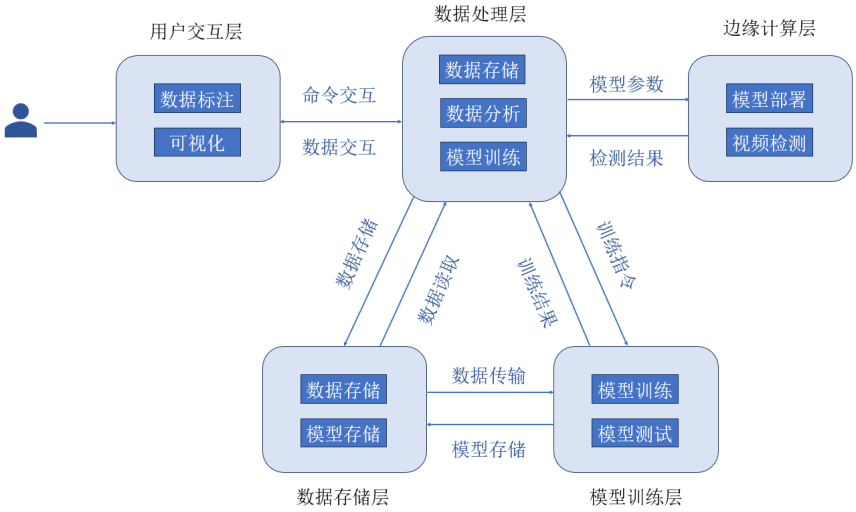


图 5-2 垃圾检测系统架构示意图

1. 用户交互层。这一层主要负责构建与用户直接交互的界面，通过前端技术框架实现。它的主要任务是提供一个直观的操作环境，以便用户能够轻松进行数据标注等活动。此外，它还负责将用户的操作请求和数据传送至数据处理层进行进一步处理。在此层中，除数据标注外的所有操作都通过数据处理层来执行，用户界面层主要起到指令传递和结果展示的作用。

2. 数据处理层。这一层作为系统架构的核心，负责协调和执行各项功能。它接收来自用户交互界面层的数据和请求，然后将任务分发给数据存储层、模型训练层和边缘计算层等。它不仅处理数据转发，还负责大部分业务逻辑的执行，包括但不限于模型的训练准备工作。鉴于其承载的通信和计算负荷较重，这一层需要配备高性能的处理器以保障系统运行的高效性。

3. 数据存储层。该层主要服务于存储数据、模型参数以及相关结果信息等。由于训练所有的数据集会持续增长，所以系统采用快速的 NoSQL 数据库作为后台支撑。同时为了提高系统的可靠性，实行了主-从备份策略，配置一台主机和两台备机，确保数据的安全和高可用性，所有存储设备均选择大容量解决方案。

4. 模型训练层。这个环境提供了模型训练和测试所需的硬件支持。在此系统

中，选用了配备大容量显存的 GPU 设备，以加速训练过程。模型的训练启动和监控由数据处理层负责，训练所需数据通过请求数据存储层获得，所有训练过程的中间结果均实时反馈至数据处理层。

5. 边缘计算层。边缘节点主要布置在户外环境，提供所需的计算资源以支持检测算法的运行。采用的目标检测算法和注意力机制结构如 Yolov8s、LADConv、GAM 等，会提前在不同类型的边缘节点上部署。模型参数由数据处理层发送，边缘节点将通过内置算法处理视频流，并将检测结果实时发送回数据处理层。

5.3 系统实现

5.3.1 开发环境

我们为垃圾检测系统构建了一个综合性的框架，融合了简洁的可视化界面、高效的数据管理机制、易于操作的神经网络训练方法及在边缘设备上的模型快速部署。本小节内容将系统地阐述不同服务器在开发阶段的环境设置及技术选型。

系统布局上，我们按功能将软硬件组合分为边缘计算节点、中心处理节点和数据存储节点。边缘计算节点涵盖了城市路面或建筑内部安装的监控摄像头以及无人驾驶智能小车等，在该系统中使用搭载 Ubuntu 20.04 操作系统的海思 HI3516DV300 开发板模拟。为了确保模型在部署时的效率与训练阶段所得结果的一致性，采用 LibTorch 作为模型推理框架。并使用 RTSP 协议以及 OpenCV 和 FFmpeg 等多媒体处理库，实现从摄像头拉取视频流的功能，再通过 ActiveMQ 消息队列技术将处理结果发送至中心处理节点。鉴于模型训练对计算资源的高需求可能会影响后端服务的正常运行，模型训练过程被安排在独立的训练服务器上。该服务器配备 4 张 Nvidia RTX 2080ti 显卡，使用 Pytorch 框架进行模型的训练。

中心处理节点的用户交互系统分为前端显示和后端逻辑处理两部分。前端部分基于 React 框架和 ElementUI 组件工具库，采用 JavaScript 进行开发。后端服务则主要基于 Python 开发，选用 Flask 框架及 Celery 工具来支撑。本系统的前后端在软硬件配置上保持较为宽松的要求，客户端操作系统支持浏览器访问

且能通过网络与服务端进行连接即可。然而，为了优化数据存取效率，建议配置较大容量的固态硬盘，以便充当数据的中转缓存。

在数据存储节点方面，数据库服务器选用 Redis 非关系型数据库，这是出于对未来数据量大幅增长时处理速度要求的考量。为了系统的稳定性，实施了主从备份策略，所有数据库服务器均采用了大容量机械硬盘。

5.3.2 模块实现

本部分内容旨在详细介绍模型训练与测试、模型维护管理以及边缘节点的配置与部署这三项关键功能，并通过泳道流程图的方式，直观呈现它们的工作原理。

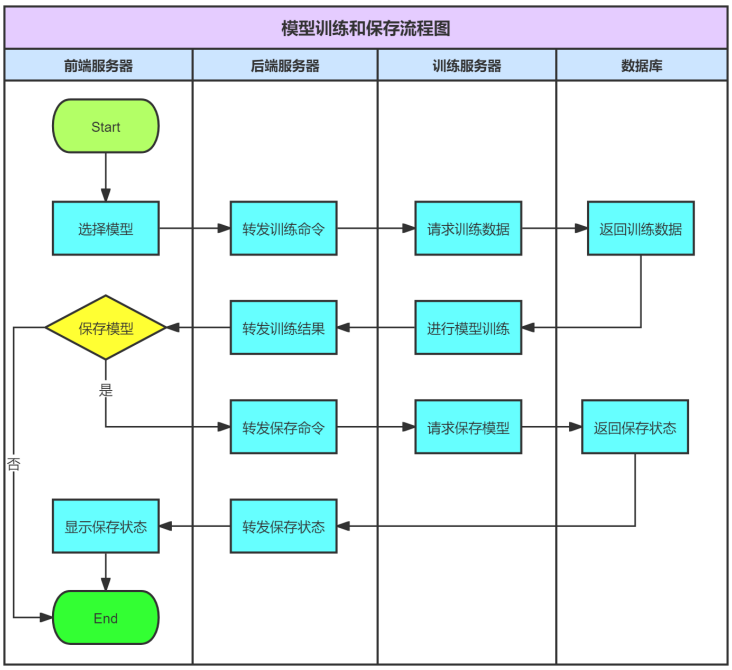


图 5-3 模型训练和保存流程图

1. 训练与测试流程：如图5-3所示流程，一旦用户发起相关模型的训练请求，后端服务器即开始整理所需的训练数据，紧接着向训练服务器发送对应的启动指令以及相关数据集编号。训练服务器据此从数据库检索相应的图像及标注信息，并依照设置的参数进行模型训练。训练过程中，通过设置周期性定时任务，不断将训练进度的更新情况反馈至前端服务器，供用户监控。待训练完成后，由用户评估模型性能，并决定是否保留该模型。

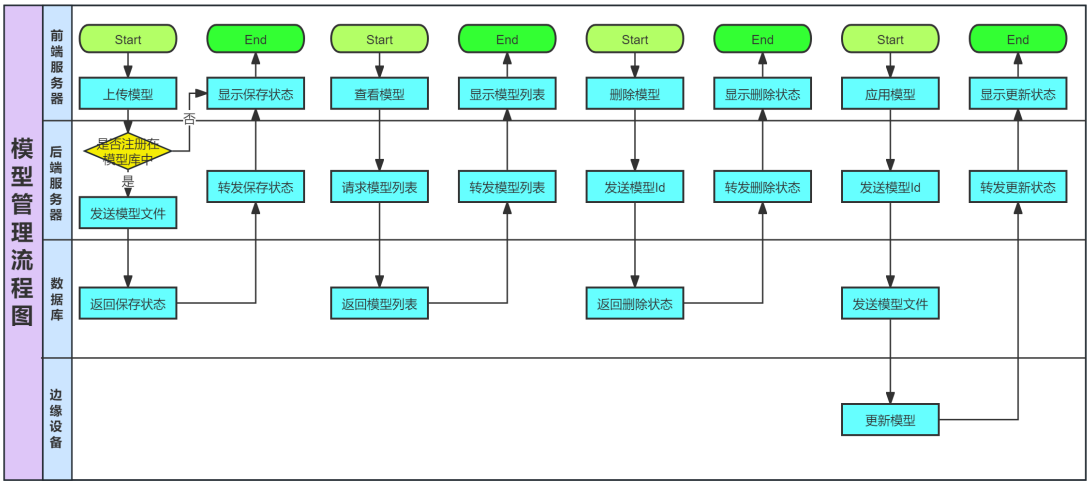


图 5-4 模型管理维护流程图

2. 模型管理维护流程：模型的管理功能模块对应流程如图5-4所示。当用户上传新的模型文件时，后端会判断该类模型是否注册在系统中，如果存在，则可手动更新模型的权重文件，若不存在则拒绝此次请求，用户可以根据需要联系开发人员更新模型库。当用户访问模型管理界面时，前端会自动请求现有模型数据，后端将此请求转发至数据库获取模型列表信息后展现给用户。用户若操作删除模型，后端将发起删除请求至数据库；若选择应用某模型，模型将直接下发至边缘节点进行参数更新，成功更新的反馈信息亦将回传前端供用户知晓。

3. 边缘节点部署：边缘节点预装有经不同组合注意力机制增强的目标检测模型及必要的支持工具，模型参数由数据库下发。边缘节点与后端服务器通过消息队列完成数据交互，确保消息的可靠传输，避免数据丢失。摄像头将捕捉到的实时视频流按 RTSP 协议送达边缘节点，利用 OpenCV 与 FFmpeg 多媒体库解析视频帧。对于无人车上部署的边缘设备，需针对每帧图像，执行目标检测任务，而对于定点摄像头搭载的边缘设备则可根据实际需要每分钟或者每数分钟检测数帧连续图像；若发现目标，则将图像及检测结果经消息队列发送至数据处理中心，以供进一步分析。

5.3.3 界面展示和使用流程

本节内容旨在展示垃圾识别系统中若干关键功能的操作界面，用户可通过浏览器进行访问和登录以体验这些功能。为确保内容的精炼与直观，我们将从

数据管理、模型管理以及监控中心三个主要模块中，分别选取一个核心功能进行详细展示。

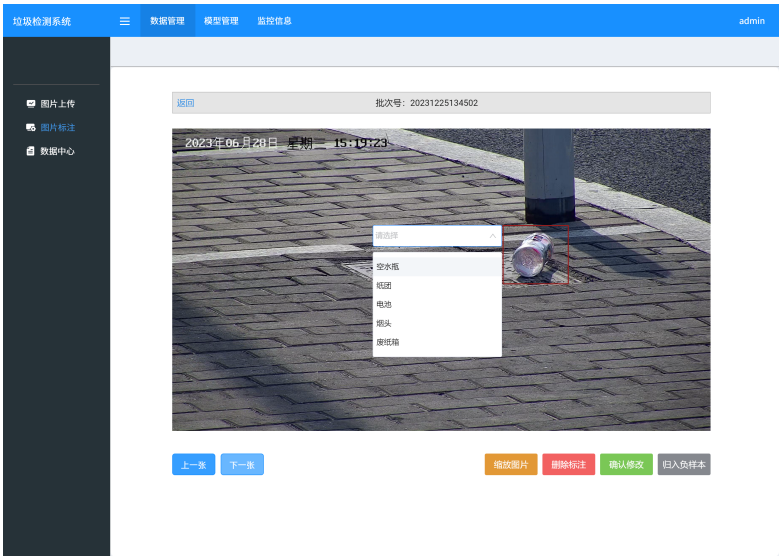


图 5-5 数据标注界面展示图

图5-5展示了垃圾检测系统中进行图片标志时的界面。观察左侧的菜单栏可知，数据管理功能被划分为图片上传、图片标注及数据中心三个主要部分。在图片上传部分，用户得以单个或批量地上传待训练的图像资料，系统随后为每次上传生成唯一的批次识别码以及相关的批次信息，涵盖了上传的时间点、图像的数量以及它们的标注状态等。点击标注按钮，界面将转向如图5-5所呈现的标注操作页，在该页顶部展示了当前处理图像的批次信息，而页面下部则布置了图片选择与功能操作相关按键。用户在图像上操作时，能够自主画出标注框，并在框绘制完毕之后通过弹出的菜单选定垃圾类别。完成所有必要的标注后，通过点击“确认修改”即可保存对该图像的标注详情。数据中心部分反馈检测出的垃圾情况，并可以按照年、月、周、日的时间维度展示垃圾的数量与种类统计情况，也允许基于特定设备名称进行数据筛选，此功能为环保部门在人力资源分配和调度上提供了极大便利，有效提升了清洁任务的执行效率。

图5-6呈现了模型管理功能区的模型训练操作界面。该区域综合了模型训练、模型管理以及参数设置三个关键子模块，其中模型管理部分支持节5.3.2所述的全面模型维护流程，包括对模型库进行添加、删除、查询，以及实现模型在边缘设备上的部署等操作。参数设置子模块存储了训练模型所需的所有参数，用户可依据需求对这些参数进行调整。在模型管理界面中点击特定模型旁的训练按

图 5-6 模型训练界面展示图

钮后，界面会转至图5-6所示的模型训练页，该页面顶部显示了与训练服务器相关的状态信息，包括服务器名称、温度、电源状态和显存占用等信息。页面中部呈现了训练过程中生成的相关损失函数图表，以便于用户对模型训练过程的实时监控。而页面底部列出了训练结束后的模型列表。用户在此可执行对相关模型的保存或删除，根据实际需求保留或清理训练成果。

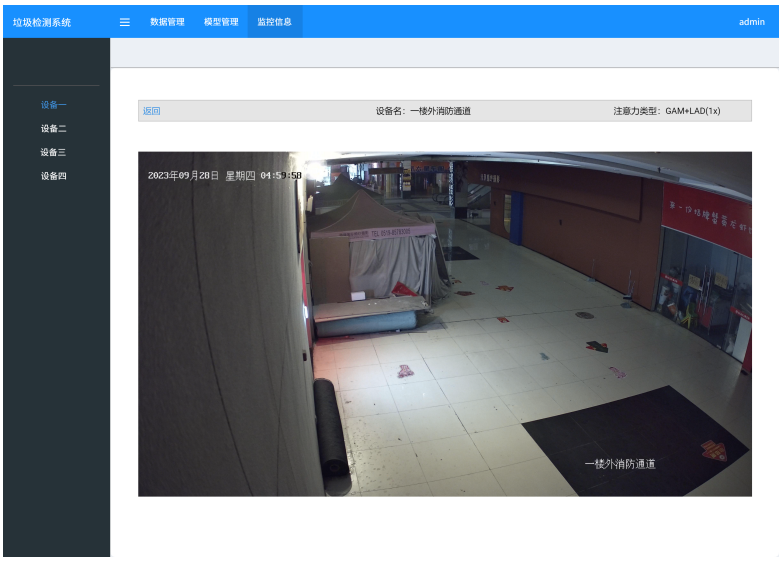


图 5-7 模型训练界面展示图

图5-7展示了各边缘设备的工作状态，该界面在左侧提供了切换不同边缘设备的功能，上方展示了对应设备的名字和具体搭载的注意力类型，而中心区域则以 RTSP 视频流的形式播放摄像头捕捉到的画面，并能够实时地反馈垃圾检测结果。

5.4 本章小结

本章详细阐述了路面垃圾检测系统的设计概念及其实现步骤。首先，我们分析了开发此系统的背景，指出了现行垃圾清理方法的局限性，并阐述了自动化检测技术所带来的显著优势。随后，通过深入分析系统需满足的核心需求，我们明确了开发目标，并据此规划了垃圾检测系统的整体框架。该框架被细分为五个主要层级，每个层级的功能、所需的开发环境以及工具等方面都进行了具体的描述。为使关键功能的工作流程更加清晰，我们通过绘制泳道图的方式对其进行详细介绍。章节最后，通过图形界面展示了系统的操作细节和实际运行情况，旨在帮助读者全面理解系统的结构及其工作机制。

第六章　总结与展望

随着卷积神经网络技术的不断进步和深化，构建复杂的网络结构成为了研究的一大挑战，而寻求高效且灵活的方法以强化现有网络架构，成为了当前研究的重要趋势。在计算机视觉任务中，引入注意力机制以促使网络更专注于图像的关键部分，已经被证明是一种有效的性能优化手段。本文从卷积核及特征图的注意力机制出发，揭示了现有研究的局限与潜在的改进空间，进而提出了两种互补的新型注意力机制。并将二者应用在具体的视觉任务中，使本文的工作不仅具有理论价值，而且实现了技术应用的实践，为满足现实需求提供了切实可行的解决方案。

在传统的卷积神经网络结构中，模型面对不同的样本仍采用一致的权重参数进行推理，这种做法忽视了输入数据多样性对特征提取精度的影响。针对这一特点，引入卷积核级别的注意力机制，可以使得网络根据输入的差异动态调整其参数，进而优化特征提取过程。目前，大多数卷积核注意力机制仍然沿用 SE 模块的思路，即通过特征通道信息获取注意力系数，并对卷积核进行加权。与此不同，本文基于卷积核捕获局部特征的独特性，采用特殊的压缩算法提取带有局部空间信息的注意力描述符，并利用自注意力机制对这些局部描述符之间的依赖关系进行学习。此外，考虑到卷积操作中不同位置的特异性，引入位置相关的注意力加权，使得网络不仅能根据输入样本的变化调整参数，而且能够根据图像中不同区域的特征重要性进行动态调整，以提炼更丰富的视觉信息。在经典数据集上的实验结果验证了本文所提出的 LADConv 卷积核注意力机制在视觉任务中相比现有技术能带来更大幅度的性能提升，同时丰富的消融实验也证实了各个模块的有效性。

在设计特征图的注意力机制时，大多数现有方法采用了一种全局视野来提取特征图的注意力。然而，随着卷积神经网络结构趋向更深层次，由此产生的特征图在通道维度上的数量不断增加，这使得全局操作在处理高维信息时可能会

导致某些关键语义信息的丢失。本文提出一种基于分组的方式来精细化地提取并应用特征图的注意力，旨在克服传统方法的局限性。通过采用结构共享和避免降维的策略，尽量降低了因分组操作而增加的额外参数和计算量。同时，引入的自注意力机制强化了分组特征间的联系，确保了在注重组内特点的同时，不失去对整体信息的把控。与其他主流的特征图注意力方法相比，本文提出的分组特征注意力机制 GAM 在性能上取得了显著的提升，并且通过与 LADConv 的结合使用，进一步发掘了卷积神经网络的潜力。最后还通过可视化的方式，直观地证明了 GAM 的优势。

除了在学术研究的相关实验验证外，本文将上述提出的两种注意力机制以不同的组合方式运用在实际的目标检测场景中。本系统融合了数据标注、模型训练及实时监控等关键环节，为目标检测任务的完整流程提供全面支持。同时特别考虑了不同边缘设备的硬件性能限制，提供了多种注意力机制的配置选项，旨在针对各种场景下的具体需求，优化现有的通用卷积神经网络模型性能。相较于传统的、一旦部署便固化配置的目标检测系统，它能让非专业人士根据需要轻松选择适合的视觉任务模型，从而在实践中更广泛地应用深度学习技术。

按照本文的现有进程，我们综合了不同领域内的学术成果，提炼出以下几条可能的改进方向，以进一步优化和深化本文的研究成果：

1. 关于卷积核的注意力机制，我们目前仍采用了通过聚集多个静态卷积核来增强模型性能的策略。尽管这种方法有效提升了性能，但同时也导致模型参数的显著增加，从而加大了模型优化的难度，影响了模型的轻量化特性。未来研究可以探索如何有效地重复利用静态卷积核[78]，减少额外参数的需求，同时保持性能的提升，以降低卷积核注意力机制实施的成本。

2. 在特征图注意力机制的设计上，目前的分组特征内部注意力提取方式缺乏足够的解释性。借鉴于数据增强领域[79-80]中的方法，如通过遮挡部分输入样本以筛选出更有效的特征点，此类策略可以被纳入特征图的注意力系数计算过程中，从而更好地实现特征重塑使网络能够关注感兴趣的特征。

3. 针对最后的目标检测系统。可以考虑更加广泛的基准网络和注意力模型，从而为用户提供更多的选择空间，以便根据具体的硬件条件和应用需求，挑选出最合适的网络配置和注意力机制组合。这种灵活性能够最大化地利用可用的硬件资源，确保目标检测系统在不同的应用场景下都能发挥最佳性能。

参考文献

[1] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[2] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12(ARTICLE): 2493-2537.

[3] KIM Y. Convolutional neural networks for sentence classification[J]. ArXiv preprint arXiv:1408.5882, 2014.

[4] CHEN Y. Convolutional neural network for sentence classification[D]. University of Waterloo, 2015.

[5] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems: vol. 25. 2012.

[6] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. ArXiv preprint arXiv:1409.1556, 2014.

[7] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.

[8] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[9] BAI S, KOLTER J Z, KOLTUN V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling[J]. ArXiv preprint arXiv:1803.01271, 2018.

[10] SALINAS D, FLUNKERT V, GASTHAUS J, et al. DeepAR: Probabilistic forecasting with autoregressive recurrent networks[J]. International Journal of Forecasting, 2020, 36(3): 1181-1191.

[11] ISMAIL FAWAZ H, FORESTIER G, WEBER J, et al. Deep learning for time series classification: a review[J]. Data Mining and Knowledge Discovery, 2019, 33(4): 917-963.

[12] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[13] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Advances in Neural Information Processing Systems: vol. 33. 2020: 1877-1901.

[14] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7132-7141.

[15] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision. 2018: 3-19.

[16] WANG Q, WU B, ZHU P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11534-11542.

[17] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems: vol. 30. 2017.

[18] DAI J, QI H, XIONG Y, et al. Deformable convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 764-773.

[19] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//European Conference on Computer Vision. 2020: 213-229.

[20] YUAN Y, HUANG L, GUO J, et al. Ocnet: Object context network for scene parsing[J]. ArXiv preprint arXiv:1809.00916, 2018.

[21] FU J, LIU J, TIAN H, et al. Dual attention network for scene segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3146-3154.

[22] GAO Z, XIE J, WANG Q, et al. Global second-order pooling convolutional networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3024-3033.

[23] XU H, ZHANG J. Aanet: Adaptive aggregation network for efficient stereo matching[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 1959-1968.

[24] MNIH V, HEESS N, GRAVES A, et al. Recurrent models of visual attention[C]//Advances in Neural Information Processing Systems: vol. 27. 2014.

[25] GREGOR K, DANIHELKA I, GRAVES A, et al. Draw: A recurrent neural network for image generation[C]//International Conference on Machine Learning. 2015: 1462-1471.

[26] XU K, BA J, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]//International Conference on Machine Learning. 2015: 2048-2057.

[27] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial transformer networks[C]//Advances in Neural Information Processing Systems: vol. 28. 2015.

[28] DAI J, QI H, XIONG Y, et al. Deformable convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 764-773.

[29] ZHU X, HU H, LIN S, et al. Deformable convnets v2: More deformable, better results[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 9308-9316.

[30] YANG B, BENDER G, LE Q V, et al. Condconv: Conditionally parameterized convolutions for efficient inference[C]//Advances in Neural Information Processing Systems: vol. 32. 2019.

[31] LI C, ZHOU A, YAO A. Omni-dimensional dynamic convolution[J]. ArXiv preprint arXiv:2209.07947, 2022.

[32] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. ArXiv preprint arXiv:1810.04805, 2018.

[33] YANG Z, DAI Z, YANG Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[C]//Advances in Neural Information Processing Systems: vol. 32. 2019.

[34] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7794-7803.

[35] LI X, ZHONG Z, WU J, et al. Expectation-maximization attention networks for semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 9167-9176.

[36] HUANG Z, WANG X, HUANG L, et al. Ccnet: Criss-cross attention for semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 603-612.

[37] GENG Z, GUO M H, CHEN H, et al. Is attention better than matrix decomposition?[J]. ArXiv preprint arXiv:2109.04553, 2021.

[38] RAMACHANDRAN P, PARMAR N, VASWANI A, et al. Stand-alone self-attention in vision models[C]//Advances in Neural Information Processing Systems: vol. 32. 2019.

[39] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. ArXiv preprint arXiv:2010.11929, 2020.

[40] YUAN L, CHEN Y, WANG T, et al. Tokens-to-token vit: Training vision transformers from scratch on ImageNet[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 558-567.

[41] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.

[42] SANDLER M, HOWARD A, ZHU M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4510-4520.

[43] LIU Z, MAO H, WU C Y, et al. A convnet for the 2020s[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 11976-11986.

[44] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images[C]//Handbook of Systemic Autoimmune Diseases: vol. 1: 4. 2009.

[45] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009: 248-255.

[46] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]//Proceedings of the European Conference on Computer Vision. 2014: 740-755.

[47] JOCHER G, CHAURASIA A, QIU J. Ultralytics YOLO[CP/OL]. 8.0.0. 2023. https://github.com/ultralytics/ultralytics.

[48] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. ArXiv preprint arXiv:2004.10934, 2020.

[49] ROSENBLATT F. The perceptron: a probabilistic model for information storage and organization in the brain.[J]. Psychological Review, 1958, 65(6): 386.

[50] KOHONEN T. The self-organizing map[J]. Proceedings of the IEEE, 1990, 78(9): 1464-1480.

[51] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[52] CHEN Y, DAI X, LIU M, et al. Dynamic convolution: Attention over convolution kernels[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11030-11039.

[53] ZHANG Y, ZHANG J, WANG Q, et al. Dynet: Dynamic convolution for accelerating convolutional neural networks[J]. ArXiv preprint arXiv:2004.10694, 2020.

[54] LI Y, CHEN Y, DAI X, et al. Revisiting dynamic convolution via matrix decomposition[J]. ArXiv preprint arXiv:2103.08756, 2021.

[55] ZHAO H, ZHANG Y, LIU S, et al. Psanet: Point-wise spatial attention network for scene parsing[C]//Proceedings of the European Conference on Computer Vision. 2018: 267-283.

[56] HENDRYCKS D, GIMPEL K. Gaussian error linear units (gelus)[J]. ArXiv preprint arXiv:1606.08415, 2016.

[57] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. ArXiv preprint arXiv:1704.04861, 2017.

[58] MAMALET F, GARCIA C. Simplifying convnets for fast learning[C]//International Conference on Artificial Neural Networks. 2012: 58-65.

[59] HOWARD A, SANDLER M, CHU G, et al. Searching for mobilenetv3[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 1314-1324.

[60] EVERINGHAM M, VAN GOOL L, WILLIAMS C K, et al. The pascal visual object classes (voc) challenge[J]. International Journal of Computer Vision, 2010, 88: 303-338.

[61] HE S, JIANG C, DONG D, et al. Sd-conv: Towards the parameter-efficiency of dynamic convolution[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 6454-6463.

[62] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533-536.

[63] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. ArXiv preprint arXiv:1503.02531, 2015.

[64] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in Neural Information Processing Systems: vol. 28. 2015.

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

[66] CHEN K, WANG J, PANG J, et al. MMDetection: Open mmlab detection toolbox and benchmark[J]. ArXiv preprint arXiv:1906.07155, 2019.

[67] XIE S, GIRSHICK R, DOLLÁR P, et al. Aggregated residual transformations for deep neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1492-1500.

[68] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4700-4708.

[69] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2980-2988.

[70] QIN Z, ZHANG P, WU F, et al. Fcanet: Frequency channel attention networks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 783-792.

[71] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]//Proceedings of the European Conference on Computer Vision. 2014: 818-833.

[72] ZAGORUYKO S, KOMODAKIS N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer[J]. ArXiv preprint arXiv:1612.03928, 2016.

[73] LEE H, KIM H E, NAM H. Srm: A style-based recalibration module for convolutional neural networks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 1854-1862.

[74] HE K, GKIOXARI G, DOLLÁR P, et al. Mask r-cnn[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2961-2969.

[75] GOYAL P, DOLLÁR P, GIRSHICK R, et al. Accurate, large minibatch sgd: Training ImageNet in 1 hour[J]. ArXiv preprint arXiv:1706.02677, 2017.

[76] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 618-626.

[77] CHATTOPADHAY A, SARKAR A, HOWLADER P, et al. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks[C]//2018 IEEE Winter Conference on Applications of Computer Vision. 2018: 839-847.

[78] LI C, YAO A. KernelWarehouse: Towards Parameter-Efficient Dynamic Convolution[J]. ArXiv preprint arXiv:2308.08361, 2023.

[79] CHEN P, LIU S, ZHAO H, et al. Gridmask data augmentation[J]. ArXiv preprint arXiv:2001.04086, 2020.

[80] YANG S, LI J, ZHANG T, et al. AdvMask: A sparse adversarial attack-based data augmentation method for image classification[J]. Pattern Recognition, 2023, 144: 109847.

致　　谢

写到这里，我即将为我过去的求学生涯画上句号。在我过往的学习经历中，研究生三年是最具压力但也收获颇丰的时光，我无数次陷入焦虑、迷茫以及自我怀疑等情绪，又无数次从消极情绪中走出寻找前进的方向。我想，也许这是每一位研究生的必经之路，也是每次进入新的人生阶段时必将拥有的情感历练。正是因为这一路上得到了诸多帮助，我才得以一次次战胜那些消极的情绪，直面生活中的种种挫折，从而深切地感受到了这三年时光的珍贵与美好。

首先，我要感谢我的导师申富饶教授。申老师不仅在学业上给予了我们莫大的支持和帮助，还时常关心学生的身心健康。即使在繁重的教学和管理工作之下，申老师仍旧坚持每周抽出时间，与我们面对面交流，耐心解答我们在科研学习中遇到的种种疑惑，提供自己宝贵的经验和见解。并且，申老师还经常分享养生之道，教导我们如何保持健康的身体，以便更高效的学习和休息。正是申老师的这种深切关怀与尊重学生的教学理念，让我们能够度过一个丰富多彩、意义非凡的研究生生活。

其次，我要感谢我的朋友们。有的在我遇到困难时提供帮助，有的在我迷茫时为我指明了方向，还有的在日常生活中给予我陪伴与支持，为我提供了丰富的情绪价值。我时常庆幸能够与你们相遇，你们无私的帮助和陪伴让我能够更轻松地面对生活的挑战，让我的人生之旅更加愉快。我衷心祝愿你们未来的道路光明璀璨，生活中充满属于你们独有的斑斓色彩。

最后，我要感谢我的家人。对我来说，家是永远的避风港。我的家人总是无条件地给予我爱与支持，我所经历的一切都离不开他们的帮助和鼓励，是他们让我能够无忧无虑地享受生活的每一个瞬间。

如果把生活比作一纸乐谱，愿我们都能毫无代价唱最幸福的歌。

简历与科研成果

基本信息

卢俣金，男，汉族，2000 年 1 月出生，四川南充人。

教育背景

2021 年 9 月—2024 年 6 月　南京大学人工智能学院　硕士

2017 年 9 月—2021 年 6 月　电子科技大学计算机科学与工程学院　本科

攻读硕士学位期间完成的学术成果

1. Junyi An, **Yujin Lu**, Lingming Zhang, Jian Zhao, Hongyuan Mei, Baile Xu, and Furao Shen. “Local-Aware Dynamic Convolution”, under-review

2. 申富饶，**卢俣金**，安俊逸，赵健，《一种基于局部特征和位置注意力的图像分类器优化方法》(202410248161.6)

攻读硕士学位期间参与的科研课题

1. 科技部重大项目“基于神经可塑性的脉冲网络高效学习机制与类脑智能系统”（参与课题年限 2021 年 9 月—2024 年 6 月），负责神经网络模型相关研究。

2. 国家电网“基于多维巡检影像匹配和对比技术的变电设备缺陷分析技术研究”（参与课题年限 2021 年 9 月—2022 年 12 月），负责图像对比与目标检测相关研究。