東南大學

电子信息学位论文

基于深度学习的野生动物识别算法研究

专 业 名 称：电子信息

研究生姓名：吕玉鑫

导 师 姓 名：路小波 教授

周鑫 高工

RESEARCH ON WILDLIFE RECOGNITION ALGORITHM BASED ON DEEP LEARNING

A Thesis Submitted to
Southeast University
For the Professional Degree of Master of Engineering

BY
LYU Yu-xin

Supervised by
Prof.LU Xiao-bo

School of Automation
Southeast University
May 2023

东南大学学位论文独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得东南大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

研究生签名：_____日期：_____

东南大学学位论文使用授权声明

东南大学、中国科学技术信息研究所、国家图书馆、《中国学术期刊（光盘版）》电子杂志社有限公司、万方数据电子出版社、北京万方数据股份有限公司有权保留本人所送交学位论文的复印件和电子文档，可以采用影印、缩印或其他复制手段保存论文。本人电子文档的内容和纸质论文的内容相一致。除在保密期内的保密论文外，允许论文被查阅和借阅，可以公布（包括以电子信息形式刊登）论文的全部内容或中、英文摘要等部分内容。论文的公布（包括以电子信息形式刊登）授权东南大学研究生院办理。

研究生签名：____导师签名：____日期：_____

摘 要

野生动物保护工作非常重要，需要及时获取动物资源种类、数量等必要信息，实施大范围的高效野生动物监测和识别。目前野生动物识别技术虽然已经有一定发展，但仍然存在数据分布不均衡、模型性能不强、技术落地应用困难等挑战。本文以通用目标识别网络YOLOX为基准，该识别模型在野生动物场景下有一定的识别准确率，但对一些困难样本的识别效果不佳，存在漏检、定位不准等问题。基于 YOLOX，本文提出了一种兼顾推理实时性和识别性能的新颖野生动物识别方法。主要工作如下：

1、针对 YOLOX 在野生动物场景下特征提取能力不足的问题，本文在第二章提出了一种基于结构重参数主干网络的野生动物识别方法。该方法基于本文设计的主干网络 BroadNet，通过 7×7 的大卷积核来增大感受野，并通过深度卷积块来减小卷积核增大带来的参数量和计算量。在训练时，使用多分支结构提高网络的表达能力，推理时转化为简洁的直筒式结构，在参数量较少、计算量较小的情况下提高了网络的感受野。该方法的整体性能表现优于 YOLOX、YOLOv4、YOLOv5 等识别方法。

2、为了解决第二章方法在部分困难野生动物样本下依然存在的定位不准、漏检等问题。本文在第三章提出了一种基于注意力机制的野生动物识别方法。该方法基于本文设计的通道-空间联合注意力模块Ultra-CBAM，在特征融合网络中选取最优位置，关注重要的前景区域，进一步提高识别效果。Ultra-CBAM 综合了 CBAM 的优势并加以优化，基于一维卷积子网络，提高了通道注意力模块跨通道的局部信息关联能力，同时使用参数化的方式优化了空间注意力聚合通道信息的过程。实验验证了本章注意力机制的有效性，第三章方法进一步提高了第二章方法的识别性能。

3、由于改进后的识别模型参数量大、复杂度高，对边缘设备不友好，本文对第三章模型进行了轻量化。本文通过使用更轻量的输入模块、深度可分离卷积以及减小特征图通道宽度等方式，降低了模型的参数量和计算量。但轻量化的过程带来了过多的性能损失，因此，本文提出了一种基于注意力引导的知识蒸馏方法来缓解这个问题。在训练轻量化模型时，本文通过注意力机制引导中间特征的蒸馏过程，生成注意力掩码对重要区域优先蒸馏。根据预测头输出生成不同位置的特征丰富度分数，优先蒸馏前景像素位置。蒸馏后的轻量化模型与其他目标识别方法相比，识别性能基本持平，但参数量和浮点数计算量大幅度减少，推理快，兼具推理效率和识别性能。

关键词：野生动物识别，注意力机制，目标检测，知识蒸馏，卷积神经网络

Abstract

Wildlife conservation is very important, and it is necessary to obtain necessary information such as the type and quantity of animal resources in a timely manner, and implement large-scale and efficient wildlife monitoring and identification. At present, although wildlife identification technology has developed to a certain extent, there are still challenges such as uneven data distribution, weak model performance, and difficult application of technology. This paper takes the general target recognition network YOLOX as the benchmark, which has a certain recognition accuracy in wild animal scenarios. But the recognition effect of some difficult samples is not good. There are problems such as missed detection and inaccurate positioning. Based on YOLOX, this paper proposes a novel wildlife identification method that takes into account the real-time inference and recognition performance. The main work is as follows:

1. Aiming at the problem of insufficient feature extraction ability of YOLOX in wildlife scenarios, this paper proposes a wildlife identification method based on structural re-parameterization backbone network in Chapter 2. In this method, the backbone network BroadNet is designed, which increases the receptive field of the neural network through a large convolution kernel of 7×7, and reduces the number of parameters and computation brought about by the increase of the convolution kernel through deep convolutional blocks. In training, the multi-branch structure is used to improve the expression ability of the network, and the inference is transformed into a concise straight structure, which improves the receptive field of the network when the number of parameters is small and the amount of computation is small. The overall performance of this method is better than that of YOLOX, YOLOv4, YOLOv5 and other recognition methods.

2. In order to solve the problems of inaccurate positioning and missing detection of the Chapter 2 method under some difficult wildlife samples. In Chapter 3, this paper proposes a wildlife identification method based on attention mechanism. Based on the channel-space joint attention mechanism module Ultra-CBAM designed in this paper, this method selects the optimal attention position in the feature fusion network and pays attention to the useful area, which further improves the recognition effect. Ultra-CBAM synthesizes the advantages of CBAM and optimizes it, based on a one-dimensional convolutional subnetwork, improves the local information correlation ability of the channel attention module across channels, and optimizes the process of spatial attention aggregation channel information by parametric method. Experiments verify the effectiveness of the attention mechanism in this chapter, and the Chapter 3 method further improves the recognition performance of the Chapter 2 method.

3. Due to the large number of parameters, high complexity, and unfriendly to edge devices, this paper lightens the model in Chapter 3. By using lighter input module, depth-separable convolution, and reducing the width of the feature map channel, the number of parameters and calculations of the model is reduced. However, the lightweight process brings too much performance loss, so this paper proposes a knowledge distillation method to alleviate this problem. The distillation process of intermediate features is guided by the attention mechanism, and attention masks are generated for priority distillation of important areas. Feature richness scores for different locations are generated based on the prediction head output, prioritizing foreground pixel locations. Compared with other target recognition methods, the distilled lightweight model has basically the same recognition performance, but the amount of parameter and floating-point number computation is greatly reduced, and the inference is fast, which has both inference efficiency and recognition performance.

Keywords: wildlife recognition, attention mechanism, object detection, knowledge distillation, convolutional neural network

目录

第一章 绪论

1.1 研究背景和意义

野生动物资源是我国重要的战略资源之一。野生动物作为生态系统的重要组成部分，与维持生态平衡与稳定息息相关。自然环境恶化、野生动物非法捕猎与交易等问题的存在，使得野生动物物种多样性锐减[1]。野生动物保护工作的基础在于及时获取动物资源种类、数量、栖息地状况等必要信息，即需要实施大范围的高效野生动物监测[2]。相比于传统人工野外调查，当前野生动物监测采用的主要方式包括红外相机、全球定位系统(GPS)项圈技术、自动感应相机陷阱技术等，具有监测方式友好、可以获取动物影像信息等优点。野外监测环境复杂，图像质量受光照影响严重，风吹草动等异动经常误触发自动感应相机, 无效监测图像比例大。监测图像中野生动物位置、大小、数量不固定，背景信息复杂。以上问题造成野生动物监测图像的人工分类劳动强度大、效率低、成本高，因此需要高效、自动化的野生动物识别技术。

在近年的野生动物识别技术研究中，最常用的方法是使用基于深度学习的目标识别方法，使用卷积神经网络进行图像特征提取，再对图像中的野生动物实现高效定位和分类。近十年，由于相关理论与计算能力的发展，深度学习技术获得了众多突破性进展，已经被广泛应用在人类生产生活中，如监控视频检测[3]、行人再识别、人脸识别等。因此可以将深度学习技术运用在野生动物监测图像识别领域，实现高效的野生动物识别。

目前野生动物图像的检测与识别技术仍然存在数据分布不均衡、模型性能不强、技术的落地应用困难等挑战[4]。在基于深度神经网络的野生动物识别方法研究中，要么是套用经典网络，要么是针对个别物种人工设计的网络。此外，野生动物识别边缘相机算力等资源紧张，部署时通常要求识别算法推理速度快、计算量小、占用显存低。因此，有必要深入研究一种高效的野生动物识别方法。该方法应该具有良好的识别性能，同时便于边缘监测设备的部署。

1.2 研究现状

1.2.1 野生动物检测识别方法

野生动物检测识别任务接受一张二维图像作为输入，对图像中存在的野生动物目标进行定位，并识别其类别。针对具体业务需要，也可以通过后处理的方式统计出野生动物目标数量。野生动物数据往往呈现长尾分布，降低了模型的识别效果。针对这个问题，蔡前舟等[5]提出了一种基于两阶段学习和重加权相结合的解决方法，并将该方法用于基于 YOLOv4-Tiny[25]的野生动物目标检测。韩家臣等[6]关注如何在单一角度和不同角度下对野生动物进行有效识别，提出一种基于 DenseNet[35]的 SSD[18]目标检测模

型，提高了对复杂目标的检测效果。杨铭伦等[7]利用YOLOv5[27]网络在红外相机等资源受限平台上实时、准确地实现海量野生动物图像自动识别，改善了野生动物监测过程中数据传输负载重、时效性低等问题。

徐其森[8]以东北虎及其食物链上的野生动物为调查背景，使用无人机搭载相机采集野生动物的行走、静止、奔跑等状态，并使用 RepVGG[44]模块对 YOLOv4-Tiny[25]进行改进，提高了该算法在野生动物场景下的检测效果。王旭等[9]认为对野生动物进行姿态追踪是保护野生动物的一种可行性方法，提出了一种基于 Transformer[43]模型的关键点检测方法，并在野生动物关键点检测的公共数据集上进行了实验，取得了良好的效果。这些方法都将野生动物识别任务作为目标检测识别技术在野生动物场景下的应用，且以深度学习技术为主。因此，接下来对基于深度学习的目标检测识别领域的研究现状进行介绍。

1.2.2 基于深度学习的目标检测识别方法

基于深度学习的目标检测识别方法同时完成对目标的定位以及分类。主要分为两类：两阶段方法和单阶段方法。

（1）两阶段方法

两阶段算法模型，以 Faster RCNN[12]等为代表，这类算法从输入图像中筛选出可能存在目标的候选区域，再利用后续的卷积神经网络完成候选区域的优化调整，得到其分类与边界框信息。Ross Girshick 等人提出 RCNN[13]（Region based Convolution Neural Network），该方法综合了传统方法和深度学习方法，使用 Selective Search[14]方法来生成类别无关的候选框。将候选框从原始图像中裁剪出来，并缩放到相同尺度，输入预训练好的 CNN 模型，对该模型进行微调，可以得到具有一定特征提取能力的 CNN 特征提取器。再使用一组 SVM 分类器对 CNN 提取的特征进行分类。

RCNN 方法有一些缺点，如训练和推理速度较慢，训练过程需要多个阶段，需要大量的存储空间等。Fast RCNN[15]在 RCNN 的基础上，设计了一个精简的训练过程来实现端到端的检测器训练，该过程同时学习分类器和边界框定位参数。在提取特征时，Fast RCNN 采用了共享特征的方式，只对原始图像进行一次特征提取，候选框的特征图可以从原始图像的特征图上进行裁剪，相比之下，RCNN 需要重复提取特征。共享特征的方式大大提高了运算效率。与 RCNN 相比，Fast RCNN 的训练速度提高了 3 倍，测试速度提高了 10 倍[16]。

Faster RCNN 在 Fast RCNN 的基础上，使用基于全卷积网络的 RPN（Region Proposal Network）网络来生成候选框。RPN 在每个特征图位置初始化 k 个不同尺度和纵横比的锚框(Anchor Boxes)。RPN 与 Fast RCNN 共享主干网络提取的特征，实现了高效的候选框提取。Faster RCNN 实际上完全由可以学习的卷积网络和全连接网络组成。

Cascade RCNN[17]结合了 RCNN 系列算法的优点，通过多个阶段的级联预测头来实现目标识别。该算法的特点是可在前一阶段检测出的候选框基础上进行进一步的筛选。

在每个阶段中，候选区域的数量都会减少，并对其中的目标区域进行更精细的预测。这种多阶段的架构使得 Cascade RCNN 具有较高的识别精度。

整体而言，两阶段算法的精度比较高，但推理速度相对较慢，并且由于经过多次候选框筛选和提取，也可能会存在召回率低的问题，对需要实时性的野生动物检测任务来说不太友好。

（2）单阶段方法

基于回归思想的单阶段算法模型，代表工作包括 SSD[18]、CenterNet[19]、CornerNet[20]、YOLO[11]系列等。这类算法不需要事先筛选候选区域，而是将最后得到的特征图划分为不同的区域，在每个区域直接输出一定数量的预测框大小以及该区域的分类信息。这类算法的精度一般略低于两阶段算法，但由于缺少了预测框的筛选和优化过程使得其检测速度更快，实时性更高。缺点是网络的训练优化比较困难，这主要是因为前景与大量的背景样本极其不均衡，Lin 等人[21]提出了 Focal Loss，为解决不平衡问题提供了新的思路。

SSD（Single Shot MultiBox Detector）可以在一个卷积神经网络中直接输出目标的类别和位置，速度和准确度均较高。SSD 是一个由卷积层和池化层构成的单阶段端到端目标识别算法。SSD 使用特征提取主干网络生成不同分辨率的特征图，这些特征图被用来识别不同大小和形状的目标。与其他目标识别算法相比，SSD 速度快、准确率高、训练简单。此外，SSD 可以在不同的场景中进行目标识别，具有一定的通用性，例如人脸检测、车辆检测、物体检测等。

FCOS (Fully Convolutional One-Stage Object Detection) [22]通过构建一个全卷积神经网络来实现目标识别。它不需要将图像分割成固定大小的小块，也不需要预先生成多个候选框，它通过构建特征金字塔来识别不同尺寸的物体，并且在每一层中都进行目标预测。此外，FCOS 还引入了中心度度量（Centerness）来提高检测效果。

YOLO 系列[11][23][24][25]目标识别算法追求最佳的速度和精度，在提出时往往汲取了可用的最先进的技巧(如用于 YOLOv2[23]的锚框，用于 YOLOv3[24]的残差网络、FPN[28]等)。YOLO 将输入图像划分为多个网格，每个网格单元负责预测中心点落在其中的物体，生成多个边界框和对应的置信度。YOLOv2 引入了 BatchNorm 层，去除了原始网络中的全连接层，并借鉴了 Region Proposal Network（RPN）的锚框机制。YOLOv3 调整了网络结构，使用残差模块构建更深的网络层，并融合不同层次的特征图来检测多尺度物体。YOLOv4 采用了丰富的优化技巧，经过筛选和算法组合，集各种最先进的训练策略，实现了效率和性能的平衡。YOLOX[26]是 YOLO 系列目标识别网络的一个优秀变种，在公开数据集 COCO[29]上表现出了优越的性能。

1.2.3 知识蒸馏与模型轻量化方法

在对野生动物进行监测时，往往需要将野生动物识别算法部署至相机等移动设备或嵌入式设备。如果识别模型参数量多、计算量大，那么就难以在这种算力较低、资

源紧张的设备上良好运行，因此对模型进行轻量化是非常有必要的。通过使用更轻量的主干网络，降低模型的网络宽度、减小模型的层数等方式，可以构建参数量为原模型十分之一乃至更小的轻量化模型。随之而来的问题是识别性能的大幅度下降。知识蒸馏为解决这个问题提供了良好的方案。

知识蒸馏是一种教师-学生(Teacher-Student)训练结构，在工业界普遍用于模型压缩。用一个预训练好的、具有较高性能的大模型作为教师模型，为学生模型的学习过程提供知识，学生模型在训练时同时接受来自教师的监督和真值数据集的监督。这种训练方式可以提升学生模型的训练效果，对构建轻量化的野生动物识别模型具有重要意义。Hinton 等人[51]认为，学生模型在知识蒸馏的过程中通过模仿教师模型输出的“暗知识”来提高泛化能力。所谓“暗知识”是教师模型训练完成后可以被用于迁移到学生模型上的各种知识形式，包括中间特征的输出、模型参数、预测结果等。本文接下来介绍两类知识蒸馏范式，一类将教师模型的最终预测输出作为知识让学生模型学习，称为基于响应的知识蒸馏[52]，另一类则关注学生模型对教师模型中间特征的学习，称为基于中间特征的知识蒸馏。

（1）基于响应的知识蒸馏

最典型的基于响应的知识称为软标签。在图像识别任务中，软标签通过“温度”系数来控制的模型的分类预测结果。软标签携带着比硬标签（如 one-hot 编码）更多的泛化信息，有助于防止学生模型过拟合。Hinton[51]将学生模型的输出和教师模型的输出分别进行软标签估计，再使用 KL 散度（Kullback Leibler divergence）损失函数进行相似度学习。Meng 等[53]认为教师模型并不一定是完美的，也可能产生错误的信息误导学生模型。为了解决这个问题，他们提出了一种条件知识蒸馏方法，当教师模型的预测响应与真值一致时，让学生模型学习教师模型的知识，否则更多地学习真值数据集。在目标检测领域，基于响应的知识除了分类头的预测结果，还包含用于定位的回归参数。Müller 等[58]让学生模型同时学习教师模型输出的分类预测结果和定位偏移预测结果，提高了对目标检测模型的蒸馏效果。

（2）基于中间特征的知识蒸馏

教师模型的预测输出包含的知识是有限的并且高度抽象。研究人员开始关注具有更加丰富知识的模型中间层，中间特征知识蒸馏可以看作教师对学生模型的提前引导或提示（Hints），可以用于解决教师和学生模型在容量之间存在的“代沟”问题。从中间层特征图中可以提取的知识类型更加丰富灵活，大大提高了传输知识的表征能力和信息量，有助于提升蒸馏训练效果。中间层特征指代神经网络中间层提取的具有一定语义信息的知识，从浅至深，中间层特征囊括了从图片的纹理、边缘、角点等低级信息到目标的形状、位置的等高级信息。

FitNets[60]通过计算一个 L2 蒸馏损失函数来比较学生模型和教师模型中间特征的差异。为了匹配教师和学生之间的语义鸿沟，Chen 等[54]提出了跨层知识蒸馏，通过注意

力机制为每个学生特征层自适应地分配合适的教师特征层进行对应引导。Kim 等[55]提出了一种因子转移蒸馏方法，通过无监督训练一个转义器来对教师模型的知识进行翻译，将翻译后的知识作为中间表征的一种更容易理解的形式，让学生模型进行学习，取得了较好的蒸馏效果。为了缩小师生之间的成绩差距，Jin 等[56]提出了一种基于路径约束优化的中间特征蒸馏方法，通过教师层引导层的输出来监督学生。Nikolaos Passalis[57]提出了一种基于概率的知识蒸馏方法，让学生模型学习教师模型中间特征在特征空间中的概率分布来进行知识转移。Huang 等[61]提出在学生模型和教师模型的特征图之间计算最大均值差异（Maximum Mean Discrepancy）作为约束，让学生模型和教师模型的中间特征图尽可能相似，让学生模型拥有接近教师模型的性能。

1.3 主要内容

本文围绕“基于深度学习的野生动物识别方法”这一实际问题，以鸨、丛林猫、黑熊、虎、雪兔、亚洲象、野马、野牛、鹦鹉、长臂猿等野生动物为研究对象，旨在提出一种能兼顾识别性能和推理效率的轻量化野生动物识别模型。

本文选取通用检测识别模型 YOLOX 作为基准模型，该模型在公开数据集上表现优秀，但在野生动物场景下，依然存在改进空间。针对 YOLOX 在野生动物场景下存在的识别不准、漏检等问题，本文通过改进主干网络及引入注意力机制等方法提高识别性能。由于改进后的野生动物识别模型参数量大、计算量大，效率低。本文设计了一种知识蒸馏方法，以较低的性能损失对模型进行了轻量化，最终提出一种兼顾识别性能和推理效率的轻量化野生动物识别模型。总而言之，本文的主要研究内容包括三部分，对应本文的二、三、四章。第二章和第三章关注如何提升性能，第四章关注如何以最小性能损失对模型进行轻量化。接下来对这几部分内容分别进行概述。

1．基于结构重参数主干网络的野生动物识别方法研究

野生动物的形态、颜色、大小等特征变化多样，场景复杂，如何提取能够反映野生动物种类的有效特征非常关键。感受野对野生动物识别任务非常重要，如果野生动物识别模型的感受野无法覆盖图像中的动物尺度，则用于识别的特征图就无法充分表达目标的语义信息，导致次优的识别效果。在第二章，本文提出了一种基于结构重参数方法和大卷积核主干网络的野生动物识别方法来提高模型的有效感受野，进而提高识别性能。

近年来的神经网络研究表明，大卷积核可以有效提高感受野。然而，基于大卷积核的网络计算量大、参数量大，训练成本高。深度卷积和结构重参数方法为这个问题提供了解决方案。结合大卷积核设计和结构重参数理论，在第二章的野生动物识别模型中，本文设计了一个具有 7×7 大卷积核的特征提取主干网络 BroadNet 来提高感受野。增大感受野后，第二章提出的野生动物识别模型可以对一些 YOLOX 漏检、识别不准的困难样本有效识别。识别效果超过了基准模型 YOLOX，也超过了其他几种高效野生动

物识别方法如 YOLOv4、YOLOv5 等。

2．基于注意力机制的野生动物识别方法研究

复杂野外环境下的动物目标易受背景信息的影响，特征提取网络能够提取的语义信息相对有限，改进后的模型对部分困难野生动物样本依旧存在定位不准、漏检等问题，需要让模型更加关注图像中具有区分性的区域。在第三章，本文提出了一种基于注意力机制的野生动物识别方法来关注图像中的前景目标。

在野生动物识别场景中，由于图像背景复杂，角度、姿态不一，注意力机制更能发挥作用。本文提出一种基于注意力机制的野生动物识别方法。在该方法中，本文设计了一种通道-空间联合注意力机制 Ultra-CBAM，Ultra-CBAM 综合了 CBAM 的优势，并加以优化创新，使之更适应野生动物场景，利用一维卷积子网络改进了通道注意力机制，通过参数化的信息聚合方式改进了空间注意力机制。基于 Ultra-CBAM，第三章方法更好地关注了图像中的前景目标，增强了特征融合过程，提高了识别用的特征图质量。在 Wildlife 测试集上，第三章方法在 mAP 和 AP50 指标上均高于第二章方法，可以有效识别一些第二章方法无法有效识别的困难样本。

3．基于知识蒸馏的野生动物识别方法研究

进行感受野增强、引入额外的注意力机制虽然提高了野生动物识别模型的效果，但增加了模型复杂度，提高了计算量，导致运行时显存和计算时间的增加。针对这个问题，在第四章本文对模型进行了轻量化。轻量化后的模型延续了第三章方法的整体设计，保留了原有结构，但大大减少了参数量和计算量。由于轻量化后的模型识别效果大幅度下降，本文进一步设计了一种知识蒸馏方法来减小识别效果的损失。

基于第三章设计的注意力机制，第四章提出一种注意力引导的混合蒸馏方法，该方法同时进行中间特征的蒸馏以及预测头的蒸馏。在中间特征蒸馏模块中，使用注意力机制生成掩码关注重要区域，在预测头蒸馏模块中，使用特征丰富度来选择更具有区分度的前景目标。轻量化模型参数量仅 3.8M，可以实时推理。蒸馏后的轻量模型与基准模型 YOLOX 相比，识别性能基本持平，但参数量和浮点数计算量大幅度减少，推理快，兼具推理效率和识别性能。

第二章 基于结构重参数主干网络的野生动物识别方法研究

2.1 概述

YOLOX 是一种高效的通用目标识别算法，它采用了多种新颖的网络设计和训练策略，能够在不同场景下进行目标识别，本文选取它作为基准模型。虽然在通用公开数据集上，YOLOX 表现出了优异的性能，但在野生动物场景下， YOLOX 的识别效果不够理想，对某些目标不全或遮挡情况下的野生动物样本无法有效识别。本文分析这可能由于其主干网络 CSPDarkNet 的特征提取效果不佳，该主干网络存在有效感受野（Effective Perceptive Field）[32]较小的问题。有效感受野的大小对于目标识别效果有重要影响[41]，基于现有文献，本章希望探索一种有效感受野更大、特征提取能力更强的野生动物识别方法。

近年来，越来越多的主干网络研究开始关注大卷积核的应用。RepLKNet[41]使用 31×31 的大卷积核，在目标识别任务上超越了 ResNext-101[34]等大型网络，达到了 Swin Transformer[43]的性能。ConvNext[42]基于 7×7 卷积在 ImageNet[10]上达到了 Top-1 的分类准确率。大卷积核参数多、计算复杂度高，往往需要通过深度卷积等方式降低复杂度，但深度卷积相比普通卷积的性能更差。这个问题可以通过结构重参数方法解决。结构重参数方法是一类可以让网络免费获得性能增益的方法。训练时，通过增加网络的宽度以及微观结构的复杂性，提高网络的容量和性能。推理时，进行模型离线转化，模型性能不变，但参数量更少。ACNet[38]是一种典型的结构重参数网络。ACNet 在训练时为 3×3 卷积增加并行的非对称卷积支路，在推理时通过参数融合，转化为单个 3×3 卷积核，融合后的主干网络结构简单、参数较少，但保留了训练时网络强大的特征提取能力。将大卷积核设计和结构重参数方法结合，允许本文以较低的代价提升有效感受野，继而提高野生动物识别模型的整体性能。

为了解决 YOLOX 在部分野生动物场景下无法有效识别目标、漏检等问题，本章提出了一种基于结构重参数主干网络的野生动物识别方法。该方法使用了本章设计的新颖的特征提取主干网络 BroadNet，采用了 7×7 的大卷积核提高感受野，通过深度卷积块来减小大卷积核所带来的计算量。本文为训练时模型设计了一种复杂的多分支结构，以保证网络的表达能力。在推理时，将其转化为简洁的直筒式结构，可以无损降低参数量、计算量以及显存占用，以较小的代价提高了感受野。增大感受野后的野生动物识别模型可以有效识别一些 YOLOX 漏检、识别不准的困难样本。在自建野生动物数据集 Wildlife 上，本章方法的识别准确率超过了 YOLOv4、YOLOv5 等其他高效通用目标识别模型。

2.2 相关工作

2.2.1 通用目标识别网络 YOLOX 在野生动物场景下的表现

YOLOX 的结构如图 2.1 所示，它采用了主干网络和头部网络相分离的设计。主干网络采用了 CSPDarknet，而头部网络则是由 YOLOv3 和 YOLOv4 的设计元素组成并加以改进的，是一种将分类头和定位头进行解耦的网络结构。

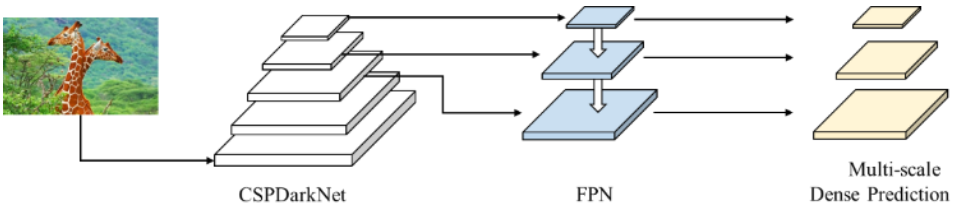


图 2.1 YOLOX 目标识别网络

虽然在通用公开数据集上，YOLOX 表现出了优异的性能，但在野生动物场景下，YOLOX 依然存在局限性。由表 2.4，YOLOX 在野生动物场景下的平均识别精准率为 41.58%，当图像比较复杂时，难以保证识别效果。如图 2.2 所示，当野生动物样本中有目标遮挡或者不全的情况时，YOLOX 存在漏检问题。

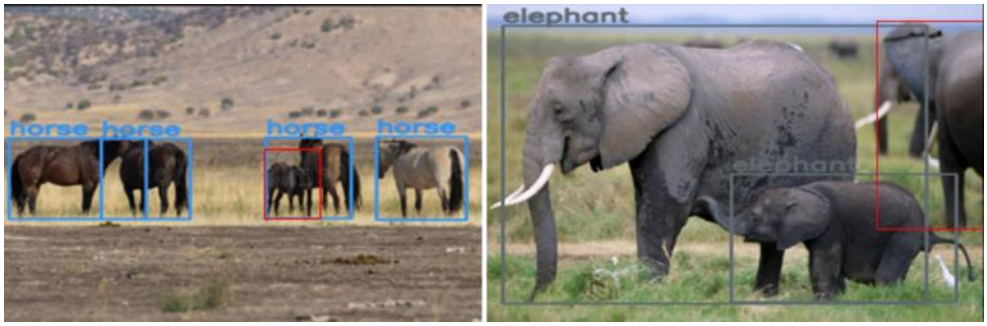


图 2.2 YOLOX 对部分样本存在漏检情况

本文将 YOLOX 在野生动物场景下识别效果不佳归因于其主干网络 CSPDarkNet。它主要由 3×3 卷积核堆叠组成，包括五个下采样阶段，通过深度叠加提高感受野。这种方式使得 CSPDarkNet 存在有效感受野较小的问题。在 2.4 节，本文对 CSPDarkNet 的有效感受野进行了可视化分析。CSPDarkNet 引入了 CSP[30]结构对网络的特征提取能力进行增强，但在复杂场景下依然可能无法提取到高质量的特征。

2.2.2 结构重参数方法与大核卷积

近年来的神经网络研究表明，增大卷积核的尺寸可以有效提高对野生动物识别任务非常重要的感受野。然而，基于大卷积核的网络计算量大、参数量大，训练成本高。深度可分离卷积和结构重参数方法为这个问题提供了解决方案。

基于卷积操作的线性可加性，结构重参数方法在训练时增加网络的宽度，提高网络模块微观结构的复杂性，增大网络的参数和容量。而在推理时进行等价转换，把复杂的网络模块转化为简单的直筒结构，模型性能不变，但参数量变少了，可以让已有

网络免费获得性能增益，也可以无损对网络进行压缩。ACNet[38]关注卷积核中不同空间位置的参数的关系以及重要程度，假设卷积核的中心十字部分权重比角落位置的权重更加重要，并提出使用非对称卷积来对普通卷积核进行增强。ACNet 为普通 3×3 卷积核并行增加一个 1×3 卷积核以及一个 3×1 卷积核，强调卷积核的中心骨干部分，增加了训练时神经网络的宽度，使得模型优化更加容易。三个卷积核的输出特征图通过直接相加的方式进行融合。由于训练时网络结构变得更加复杂，参数空间更大，模型的表达能力得到了提升。而推理时，经过等价转化后的模型并没有增加任何复杂性，也不需要任何额外的计算资源，既提升了网络的性能，又没有损失推理效率。Diverse Branch Block[39]将 ACNet 进一步扩展和改进，提出了更加复杂的并行结构，探讨了对现有网络结构重参数的几种方法。RepVGG[40]沿用结构重参数思路，拓宽了经典模型 VGGNet 的性能边界，在 ImageNet 上的 top-1 准确率超过了 80%。

丁霄汉在 RepLKNet[41]中思考了卷积神经网络结构设计中大卷积核的应用，并基于深度可分离卷积实现了具有 31×31 超大卷积核的主干网络。该网络浅而宽，应用了深度可分离卷积以及重参数结构来构建大核模块，具有较大的有效感受野。近期工作中，Swin Transformer[43]使用滑动窗口的方式来捕捉空间特征，其窗口大小可达到 12，直接提升网络的有效感受野，也可以看作是大卷积核的一种变种。Liu 等人[42]使用 7×7 的深度可分离卷积设计了性能强大的网络 ConvNext。在 ImageNet 上，ConvNext 与 Swin Transformer 性能相当。[44]进一步将卷积核大小扩展到 51×51，其性能也与最先进的分层 Transformer 相当，证明了大卷积核设计的性能价值。

根据以上分析，本文接下来提出一种基于结构重参数主干网络的野生动物识别方法。该方法规避了 YOLOX 识别模型的缺点，创新性地利用了大卷积核的感受野增益能力以及重参数方法的模型增强能力，以 7×7 卷积实现了一种具有较大有效感受野的主干网络 BroadNet，实验表明，基于该主干网络，本章方法可以解决部分困难野生动物样本的识别问题。

2.3 基于结构重参数主干网络的野生动物识别方法

为了解决 YOLOX 在部分野生动物场景下的漏检、识别不准等问题，本节提出一种基于结构重参数主干网络的野生动物识别方法。该方法通过深度卷积以及大卷积核增大了识别模型的有效感受野，通过结构重参数方法，设计了一种性能强大的多分支多尺度特征提取结构，提高了网络的表达能力。接下来本文先在整体上对本章提出的野生动物识别方法进行概述，再介绍本章方法设计的新颖主干网络 BroadNet 以及多分支模块 Broad Block。

2.3.1 整体设计

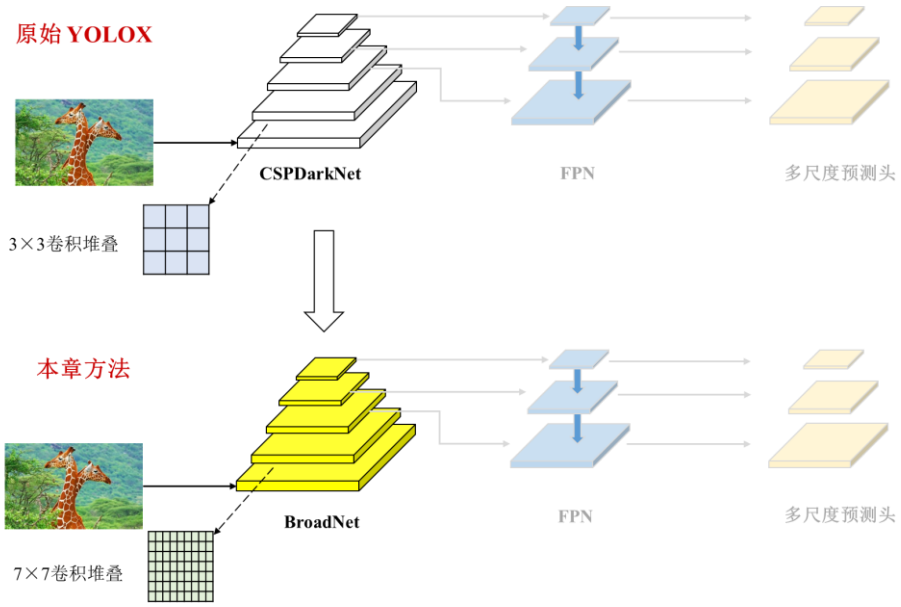


图 2.3 基于结构重参数主干网络的野生动物识别方法

本章方法和基准方法 YOLOX 的对比如图 2.3 所示，只需用本章设计的强大主干网络 BroadNet 替换原有主干网络便可以取得不错的性能提升。输入野生动物图像样本经过预处理后送入本章设计的主干网络 BroadNet 中，经过 5 个下采样阶段以及多个特征提取模块，提取出具有不同层次语义信息、不同分辨率的特征图。最后三个层级的特征图 C3、C4、C5 分别送入特征融合网络中进行特征融合，并分别使用一个预测头进行前景目标的定位和分类。

BroadNet 综合了 Inception 系列网络与结构重参数方法的优势，使用 7×7 大卷积核来增大整体野生动物识别模型的有效感受野。由于大卷积核带来了更多的参数量和计算量，本章基于深度可分离卷积改进出了参数量更少的深度卷积块结构，基于该结构实现 7×7 卷积，可以大大降低参数量和计算量。构成 BroadNet 的主要特征提取模块 Broad Block 具有可缩放网络宽度的扩张层以及多分支的大卷积核特征提取块，可以进行多个尺度的特征提取和融合。

相比于同类野生动物识别模型，本章方法对主干网络特征提取模块的基本逻辑、卷积核的实现方法、多分支结构的融合方法都进行了改进和创新，更加适合野生动物识别场景。接下来分别对本章方法的具体设计进行阐述。

2.3.2 主干网络 BroadNet 设计

BroadNet 综合了 Inception[35]系列网络的结构优势，通过多分支结构提取不同层次的特征，丰富了特征多样性。基于结构重参数方法，BroadNet 规避了其推理时显存占用过大的劣势，针对野生动物识别的场景特点，使用 7×7 的大卷积核以及深度卷积等技术提取特征，以较低的参数量和计算代价获取了较大的有效感受野。

用于目标识别任务的主干网络通常包括 5 个下采样阶段，随着网络深度的增加，输

出特征图的通道数随之增加，且具有一定的冗余性。因此需要对特征图进行下采样来减少显存占用和计算耗时。BroadNet 也由一个输入模块 stem 和 4 个阶段的下采样层和特征提取模块组成，结构概况可见图 2.4（a）。表 2.1 给出了 BroadNet 的详细网络参数。BroadNet 共有 20 层参数层，相比于 CSPDarkNet、ResNet 这类动辄四五十层乃至上百层的主干网络比较小巧。

输入模块是网络的开始部分，它接受原始图片作为输入。为了让下游的分类器和定位器具有良好的表现性能，需要在主干网络的开始部分就获取到丰富、冗余的低维图像特征。BroadNet 的输入模块延续了 CSPDarkNet 的设计，如图 2.4（b）所示。3×3 卷积对原始图像进行初步处理，捕捉粗粒度的特征信息。stride=2 的 3×3 卷积对特征图进行下采样，降低了网络后续阶段的运算量。

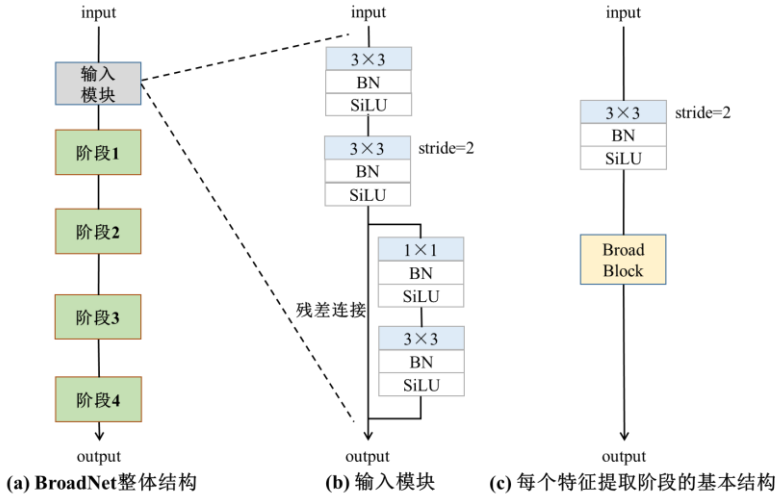


图 2.4 BroadNet 主干网络结构

Stage 1-4 是网络的主要部分，用于提取不同语义层次的图像特征。如图 2.4（c），每一个阶段中，先对输入特征图进行下采样。后续的 Broad Block 从下采样后的特征图中提取更高层次的特征。Broad Block 具有可缩放网络宽度的扩张层和多尺度多分支的大卷积核特征提取块，以及用于特征选择的收尾层。接下来对 Broad Block 进行详细介绍。

表 2.1 基于 7×7 大卷积核的 BroadNet 网络参数

Stage	Type	Filters	Kernel size	Output size
Stem	卷积块	32	3×3	416×416
	下采样卷积块	64	3×3, stride=2	208×208
	Res Block[32]	64	3	208×208
Stage 1	下采样卷积块	128	3×3, stride=2	104×104
	扩张层	256	1×1	
	多分支结构	256	7	
	收尾层	128	3×3	

表 2.1（续）

Stage	Type	Filters	Kernel size	Output size
Stage 2	下采样卷积块	256	3×3, stride=2	52×52
	扩张层	512	1×1	
	多分支结构	512	7	
	收尾层	256	3×3	
Stage 3	下采样卷积块	512	3×3, stride=2	26×26
	扩张层	1024	1×1	
	多分支结构	1024	7	
	收尾层	512	3×3	
Stage 4	下采样卷积块	512	3×3, stride=2	13×13
	扩张层	1024	1×1	
	多分支结构	1024	7	
	收尾层	512	3×3	

2.3.3 多分支大卷积核模块 Broad Block 设计

Broad Block 是 BroadNet 的主要组成部分，占据了大部分参数。它的训练时和推理时结构不同，可以互相转化。在训练时是一种基于深度卷积块的多分支大卷积核模块，在推理时则将多分支结构转化为直筒式结构，降低参数量。在不同下采样阶段堆叠 Broad Block，可以有效提高网络的感受野，增强特征提取能力。

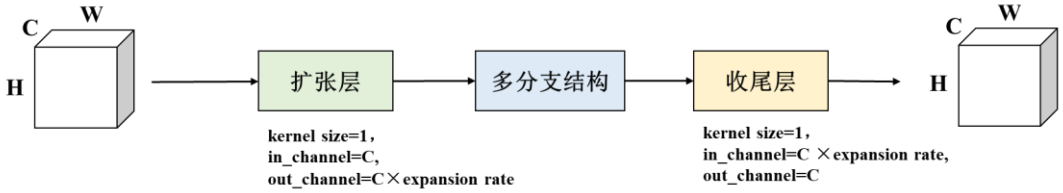


图 2.5 Broad Block 训练时结构

Broad Block 的训练时结构如图 2.5 所示，每个 Broad Block 包括三部分：扩张层（expansion layer），一个具有大卷积核的多分支特征提取结构，以及收尾层（Tail layer）。扩张层是每个 Broad Block 与前一个阶段的接口层，它包括一个 1×1 卷积核，一个 Batch Norm 层，一个 SiLU 非线性函数。它的作用是通过一个扩张因子参数 expansion rate 以一定比例增大本层的特征图通道数，从而控制当前 Broad Block 的网络容量和宽度。SiLU 函数的引入为网络提供非线性表达能力。通过 expansion rate 对网络宽度进行缩放，使得 Broad Block 是一个中间大、两头小的纺锤形结构。在 2.4 节，本文将对 expansion rate 对网络性能的影响进行讨论。收尾层与扩张层类似，使用 1×1 卷积、BatchNorm 层和 SiLU 非线性函数实现。它的作用是在多分支结构输出的特征图基础上进行特征选择和通道压缩。

多分支结构是主要的特征提取结构，具有 7×7 大卷积核，可以提取并融合多个层次的特征图。它接受具有多通道冗余信息的高维特征图输入，保证了网络的表达能力。

接下来对它进行详细阐述。

（1）训练时多分支特征提取结构

多分支结构是 Broad Block 中的主要特征提取模块。它主要由 7×7 的大卷积核构成，而非一般主干网络里常用的 3×3 卷积。一个 7×7 卷积核的感受野与 3 个 3×3 卷积核的感受野相同，这使得 BroadNet 能够以较小的深度达到较大的感受野。这同时避免了深度增加可能带来的梯度消失问题。

如图 2.6 所示，训练时的多分枝结构包括一个 7×7 卷积路径，一个 1×1 卷积路径，一对非对称卷积（7×3，3×7）路径，以及一个 7×7 平均池化路径。每条路径中的卷积核后均使用 BatchNorm 层进行标准化。3×7 路径和 7×3 路径有助于对 7×7 卷积的中心十字部分进行增强。此外，通过这种非对称的卷积方式，在水平方向和竖直方向分别提取多尺度的特征，进而可以在主干网络的层级就实现多尺度的特征融合。在设计时，本文将平均池化看作一种恒等连接，它以等比例强制保留卷积窗口中各个位置的信息，通过本章的消融实验，证明平均池化对最终模型的性能具有重要作用。1×1 卷积路径则给予了网络在通道层面更大的缩放性。

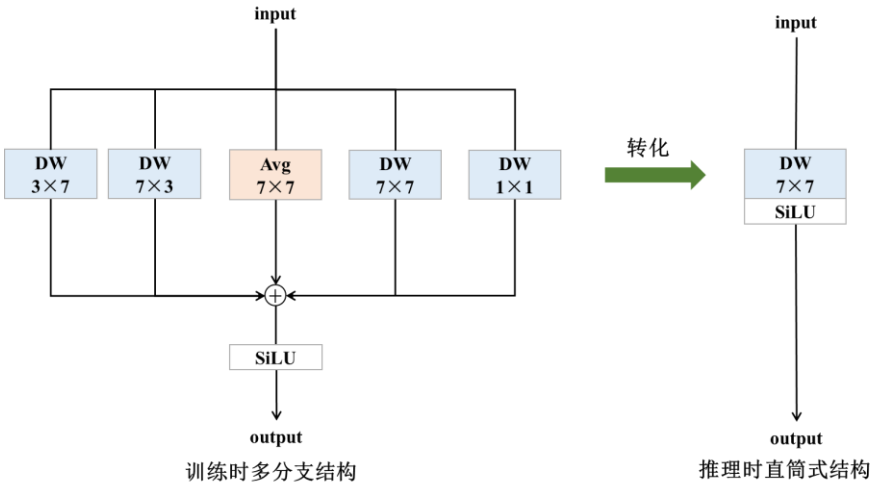


图 2.6 Broad Block 模块的多分支结构

在推理时，这 5 条卷积路径通过一定的融合方法转化为单个直筒式 7×7 卷积，降低了推理时的参数量、计算量以及显存占用。不同的融合方法会对融合效果产生影响，接下来对本文设计的多分支结构含参融合方法进行阐述。

（2）多分支结构的含参融合方法

多分支结构中的各个支路共享输入特征图，分别进行特征提取工作，再融合为一个输出特征图。融合方法的选择对最终性能具有重要的影响。本文考虑了多种融合方法，这里介绍一种无参的简单相加方法，以及四种含参融合方法。

简单相加的方法将各支路的输出直接相加。实际上假定每个支路的输出具有相同的权重比例，是一种无参的融合方法。考虑到各个支路的贡献可能是不相等的，有必

要讨论几种隐式的、自适应参数化权重的融合方法，为各个支路增加乘加权重，在网络的训练过程中由网络自己决定如何分配。基于这一思路，本文进一步考虑四种参数化权重融合的方法，分别为基于 SoftMax 归一化的融合、基于 ReLU 归一化的融合、基于绝对值归一化的融合，以及无数值限制的融合。为 5 个支路进行编号，对支路 i，令其输出特征图为 F_i，乘加权重为 W_i，则融合后的输出特征图可统一表示为：

$$Output = \sum_{i=1}^{5}\left(F_i \cdot W_i\right) \tag{2.1}$$

接下来分别对几种有参方法进行介绍。

基于 SoftMax 归一化(SoftMax-based)的融合。直觉上，SoftMax 输出一个[0, 1]内的概率值，可以方便地为每个支路给予权重。在融合时，应用 SoftMax 函数对 w_i 进行归一化，将其值域变换为[0, 1]范围内，代表每个支路的重要程度。这种情况下，有：

$$W_i = \frac{e^{w_i}}{\sum_{j=1}^{5} e^{w_j}} \tag{2.2}$$

基于 ReLU 归一化(ReLU-based)的融合。SoftMax 函数计算过程涉及到比较多的指数运算，基于 ReLU 的归一化方法使用 ReLU 函数对 w_i 进行处理，计算量更小。ReLU 将负值输入置零，正值和 0 保持不变。这种方法的考虑下，认为权重小于 0 的分支对网络贡献不大。乘加权重可表示为：

$$W_i = \frac{ReLU\left(w_i\right)}{\sum_{j=1}^{5} ReLU\left(w_j\right)} \tag{2.3}$$

基于绝对值归一化(Abs-based)的融合。绝对值归一化的融合方法将权重小于 0 的分支也考虑到，认为绝对值幅值更大的分支具有更大的贡献，乘加权重可表示为：

$$W_i = \frac{Abs\left(w_i\right)}{\sum_{j=1}^{5} Abs\left(w_j\right)} \tag{2.4}$$

无数值限制（No-Bound）的融合。这种方法为每个支路简单地乘上一个可学习权重。有 $W_i = w_i$。其实质是对每条支路的输出数值进行简单缩放，这种方式不需要对权重进行额外的运算，但实验表明其效果相比于其他几种方式是最差的。由于权重在优化时不受限制，可能会使得网络的训练过程不稳定，导致次优的结果。

本文通过消融实验确定了最终采用的融合方法，可见 2.4 节。简言之，含参的融合方法对 BroadNet 的性能具有重要作用，基于绝对值归一化的融合方法效果最好。

（3）基于深度卷积块的大卷积核设计优化

在多分支卷积结构中使用了 7×7 卷积来提取特征，大卷积核会带来计算量以及参数量的增加，这是很多网络不使用大卷积核的原因。对这个问题，本章的解决方法是使用深度卷积块（Depthwise Convolution Blocks, DW）来实现 7×7 卷积。深度卷积块

基于深度可分离卷积[63]，可以大幅度降低大卷积核所需要的计算量和参数量。为了推理时融合的需要，其他支路的含参卷积核也使用深度卷积块来实现。

深度可分离卷积块包括依次相连的深度（depthwise）卷积和逐点（pointwise）卷积，如图 2.7。以 44×44×3 的输入特征图，7×7 的卷积核为例，深度卷积的作用是提取空间特征，关注局部空间位置的信息，对输入特征图进行分组卷积，每个通道均作为独立一组，卷积核参数张量的通道数与输入特征图相同。卷积核的每个通道与特征图的每个通道是一一对应的关系。深度卷积只关注特征图的空间信息，没有通道间信息交互，不进行通道维度变换。逐点卷积使用普通 1×1 卷积来完成这些必要的功能。

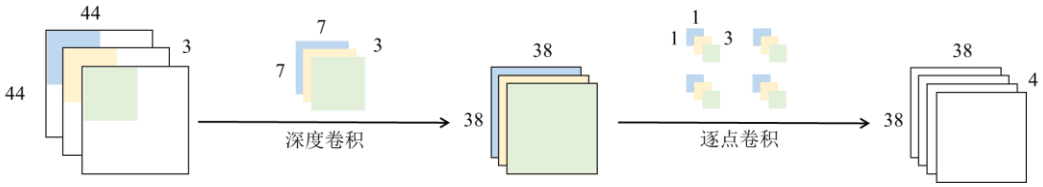


图 2.7 深度可分离卷积块

与之不同，深度卷积块包括依次相连的一个 7×7 深度卷积和一个 1×1 深度卷积，如图 2.8 所示。由于 Broad Block 多分支结构后的收尾层可以完成通道间的信息交互，可以将 1×1 卷积也进行深度分组化，仅仅对特征图的每个通道分别进行线性映射，这样可以节省一点计算量。

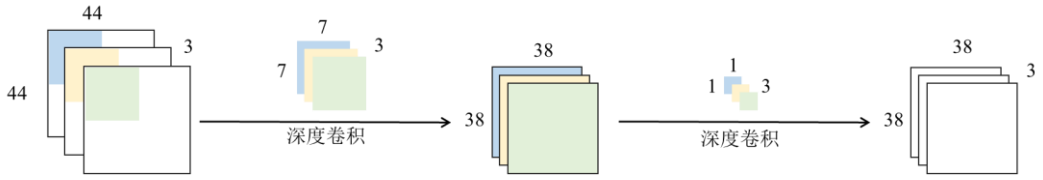


图 2.8 深度卷积块

（4）深度卷积块的推理时转化

不管是 7×7 深度卷积还是 1×1 深度卷积，本质上都是分组数等于输入通道数的分组卷积。由于 1×1 深度卷积只对输入的每个通道进行线性映射，不涉及空间特征的提取，因此可以将 1×1 深度卷积的每个通道与 7×7 卷积对应融合，如图 2.9 所示。以通道数为 3 的特征图为例，输入特征图被分为 3 组，即图 2.9 中的虚线部分。每组的特征图依次经过 7×7 深度卷积提取空间信息，再通过 1×1 卷积映射。通过以下步骤可以将深度卷积块转化为一个分组卷积核：

1. 分别把 7×7 卷积块、1×1 卷积块中的卷积层和 BatchNorm 层融合为普通分组卷积层,依次相连的卷积层和 BatchNorm 层的融合过程可参考文献[39]。

2. 将依次连接的 1×1 卷积和 7×7 卷积进行分组间的一一对应。

3. 在每组内将 1×1 卷积和 7×7 卷积对应参数相乘，得到该组融合后的 7×7 卷积参数张量。

4. 将每组的7×7卷积参数张量在通道维度进行拼接，得到融合后的完整卷积核参数，该卷积核可以用于代替初始的深度卷积块，如图 2.9（b）所示。

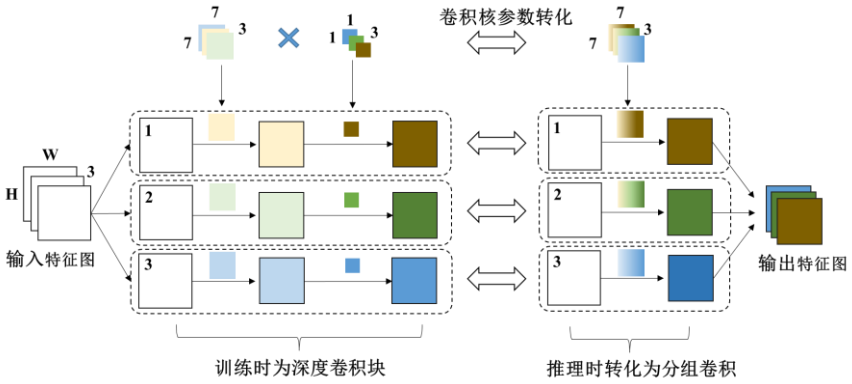


图 2.9 深度卷积块融合过程（以 channel=3 为例）

（5）多分支结构的推理时转化

一个具有 C 个输入通道、O 个输出通道、卷积核大小为 $K\times K$ 的卷积层的参数可以表示为具有 4 个维度的张量 $T\in\mathbb{R}^{C\times O\times K\times K}$。除此之外，通常还有偏置参数 $b\in\mathbb{R}^{O}$。假定输入特征图为 $F\in\mathbb{R}^{B\times C\times H\times W}$，其中 B 表示训练时的 Batch size。输出特征图表示为 $F'\in\mathbb{R}^{B\times O\times H'\times W'}$，其中 H' 和 W' 表示输出特征图的高和宽。用 $\circledast$ 来表示卷积运算，取输出特征图 F' 上第 i 通道位置的 (j,k) 点，该点的值可形式化表示为

$$F'_{i,j,k}=\sum_{c=1}^{C}\sum_{m=1}^{K}\sum_{n=1}^{K}T_{c,i,m,n}S(j,k)_{c,m,n}+b_j \tag{2.5}$$

$S(j,k)$ 表示在空间像素坐标 (j,k) 处的 $K\times K$ 滑动窗口内的特征子图。输出特征图 F' 上任意一点由卷积核参数与特征图上对应位置处特征值通过线性乘加运算得到，从而有

$$a(F\circledast T)=F\circledast(aT) \tag{2.6}$$

用 $T^{(1)}$ 和 $T^{(2)}$ 表示两个独立的卷积核参数张量。当两个卷积核具有相同的 padding、stride、kernel size、output channel 等参数时，在同一空间位置 (j,k)，两个卷积核与相同的特征子图相对应。如图 2.10 所示。

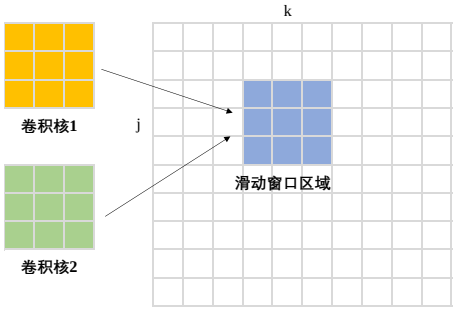


图 2.10 在同一位置处，两个 K×K 卷积核对应特征图上相同的滑动窗口

对于 $F \circledast T^{(1)} + F \circledast T^{(2)}$，由于两个卷积核的作用于相同的特征图位置，因此有

$$F \circledast T^{(1)} + F \circledast T^{(2)} = F \circledast \left(T^{(1)} + T^{(2)} \right) \tag{2.7}$$

根据式 2.7，当把输出特征图进行相加时，我们可以融合两个并行的 K×K 卷积层分支，如图 2.11。令 $T' = T^{(1)} + T^{(2)}$，$b' = b^{(1)} + b^{(2)}$，以 T' 和 b' 就可以构建一个等效卷积核替代原始分支。当其中一个卷积层为平均池化层时，可将其看作参数均为 $1/(K \times K)$ 的普通卷积进行融合。

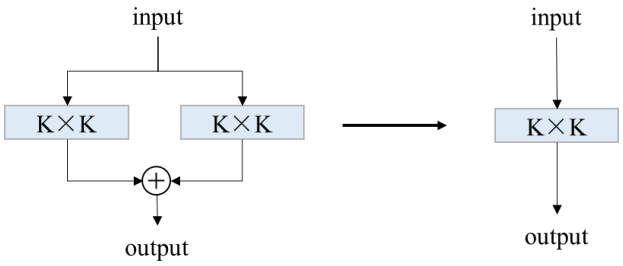


图 2.11 利用卷积操作的可加性融合两个并行卷积分支

当两个并行分支的卷积核大小不同时，以 1×1 卷积核和 3×3 卷积核为例，可将 1×1 卷积核视为外圈参数为零的 3×3 卷积核，如图 2.12 所示。这种等价看待基于相同的 stride、相同的输出通道数，对于 3×3 卷积核，需要在特征图上进行 padding 以保证一致性。

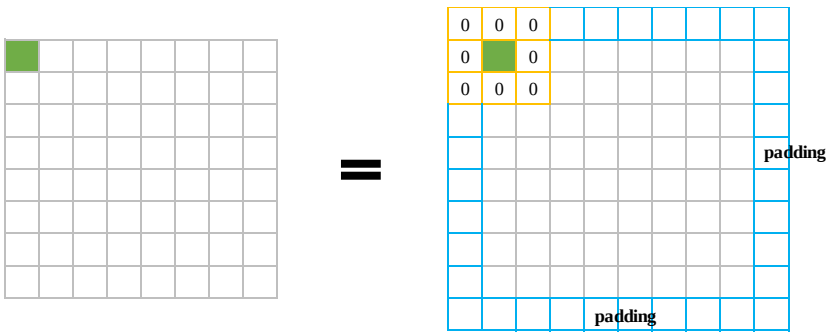


图 2.12 1×1 卷积等价于 padding=1，外圈参数为零的 3×3 卷积

同理，多分支的每个支路，经过对齐后，可视作等价 7×7 卷积核，这些卷积核的同一空间位置相互对应。这里形成对应关系，有两个前提条件，一是对卷积核进行合理 padding，以保证核与核之间的中心对齐；二是所有支路均等价于通道数相同的分组卷积，保证同一通道位置可以进行一一对应，如图 2.13。整体转化过程可以描述为：

1. 将路径中所有的深度卷积块依据前文所述方法转化为等价分组卷积核。
2. 将平均池化层看作尺度同等大小的带参卷积核。
3. 将每条支路的卷积核参数张量进行融合，得到等价卷积核参数张量。
4. 构建一个新的分组卷积核，并将 3 中得到的等价参数张量赋给该卷积核。

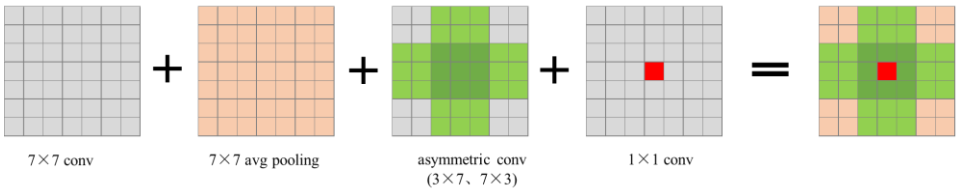


图 2.13 Broad Block 的分支结构转化为单个卷积块过程

基于上述转化过程，可将 Broad Block 训练时的多分支结构转化为一个普通的 K×K 分组卷积，从而降低推理时网络的时间复杂度和空间复杂度。将 BroadNet 中每个 Broad Block 都进行转化，就得到了 BroadNet 的推理时网络。

训练时和推理时网络的主要差距体现在运行时空间的占用上，由于推理时不存在多分枝结构，因此网络在运算时不需要保存过多的中间结果，减轻了对显存的需求。表 2.2 给出了训练时网络和推理时网络的参数量（Params/MB）、浮点数计算量（FLOPs/G）以及运行时内存的消耗（Memory/M）。在不损失性能的前提下，推理时转化减少了网络的参数量、浮点数计算量，大大减少了运行时内存的消耗。推理时网络的运行空间占用大小约降低 50%。随着网络宽度系数 expansion rate 的提高，这个差距还将进一步增大，由此可见推理时转化是必要的。

表 2.2 BroadNet 训练时网络和推理时网络参数量、计算量、运行空间占用对比

Model	Expansion rate	Params (M)	FLOPs (G)	Memory (M)
BroadNet-train	1.0	5.32	12.31	362.76
BroadNet-val		5.24	11.96	247.89
BroadNet-train	2.0	6.68	15.23	525.48
BroadNet-val		6.52	14.53	295.75
BroadNet-train	3.0	8.05	18.15	688.21
BroadNet-val		7.81	17.10	343.61
BroadNet-train	4.0	9.41	21.07	850.94
BroadNet-val		9.09	19.67	391.47

2.4 实验与分析

本节基于实验验证本章方法的有效性，并与其他野生动物识别方法进行对比。本文中所有实验均基于同样的数据集、实验设置下进行，以保证对比的公平性、准确性。接下来先对所用的野生动物数据集、评价指标、实验设置等进行介绍，再介绍各种消融实验以及与 BroadNet 与其他主干网络的对比、本章方法与基准网络等对比实验结果。

2.4.1 Wildlife 数据集介绍

野生动物场景下的公开数据集资源较少，我们使用自己构建的 Wildlife 数据集作为实验基准数据集，数据集由人工标注，图片来自于互联网。一些样本如图 2.14 所示。野生动物场景下的图像具有背景复杂、目标大小不一等特点，部分样本存在目标遮挡、不全等问题，这为精确识别相关动物目标增加了难度。

图 2.14 Wildlife 野生动物数据集

本文按照 7：1 的比例在每个类别进行随机采样，划分出训练集和测试集。训练集包括 3680 张图片，测试集有 525 张图片。这些图片分布在 10 个类别，包括丛林猫，亚洲象，老虎，野牛，野马，长臂猿，兔，鸨，鹦鹉和黑熊。整体数据集样本比例呈现不平衡分布，如图 2.15 所示。数量最多的虎类的数量几乎为最少的猫类的 10 倍。

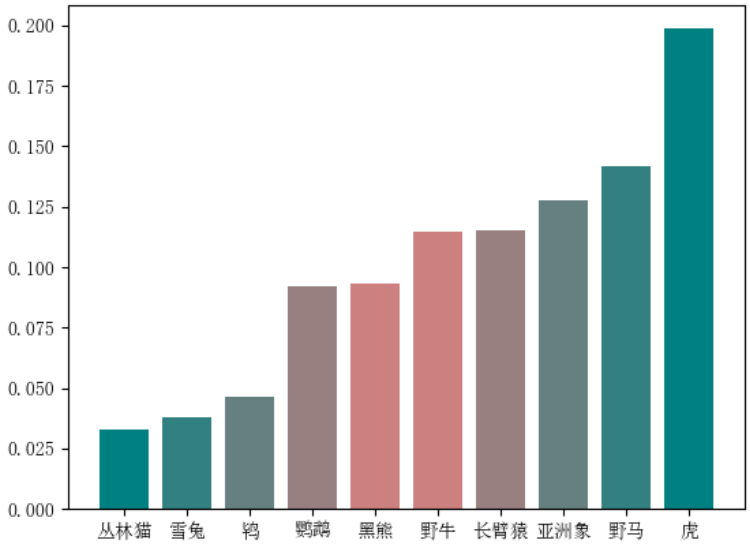


图 2.15 Wildlife 野生动物数据集各类别比例

2.4.2 评价指标

本文实验以各类别的平均精准率（mean Average Precision，mAP）作为性能评价指标。根据正样本定义的不同，本文给出 mAP0.5（文献中也称 AP50）和 mAP 两种指标。前者表示检测框与真值框 IOU 大于阈值 0.5 时作为正样本，是一种较为粗糙的指标，适用于对定位精度要求不高，更加关注检出率的情况。后者则取 COCO 数据集中的标准，从 0.5-0.95 每隔 0.05 取一个阈值计算 mAP 值，最后计算各阈值下的平均值，可以反应模型的综合水平。mAP 通过召回率（Recall）和精准率（Precision）计算。

Recall (召回率): Recall 定义为正样本被正确检测到的比例，通常以 True Positive (TP) 数量除以真实的正样本数量来计算。可以用式 2.8 来表示：

$$Recall = \frac{TP}{TP + FN} \tag{2.8}$$

其中 TP 是正确检测到的正样本数量，FN 是误判为负样本的正样本数量。

Precision (精确率): Precision 定义为正确检测到的正样本数量与检测出的所有正样本数量之比，通常以 TP 数量除以被检测出的正样本数量（包括正确检测的正样本和误判的正样本）来计算。可以用式 2.9 来表示：

$$Precision = \frac{TP}{TP + FP} \tag{2.9}$$

其中 FP 是误判为正样本的负样本数量。

Mean Average Precision (mAP): mAP 是评估目标识别算法在不同精确率-召回率下的性能的一种指标。mAP 通常以精确率-召回率曲线下的面积来计算，mAP 能够综合反映出算法在不同召回率下的精确率水平，从而更好地评估目标识别算法的效果。具体地，AP 可以通过以下公式表达：

$$AP = \sum_{n=1}^{N} \big(R(n) - R(n-1)\big) \times P(n) \tag{2.10}$$

其中 P(n)是当召回率为 R(n)时的精确率。正样本是否正确检出，采用 IoU 来计算。它表示预测框与真值框之间的交并比。IoU 取值范围在[0, 1]之间，其中 1 表示预测框与真实框完全重合，0 则表示两个框不相交。

2.4.3 实验基础设置

本文实验基于 ubuntu20.04，在两张 RTX 3060GPU 上进行并行训练，在单张 3060 上进行推理。在模型训练时，为了控制变量和保证比较效果不受其他如随机性等因素的影响，本文固定了随机数种子，并保持无关超参数一致。所有实验采用了标准数据增强，包括 HSV、随机翻转、MixUp、归一化等。训练过程未使用预训练模型，所有实验均从头进行训练，设置 Batch size=16，在 100 个 epoch 内达到收敛状态。在训练的初始阶段，为了避免初始学习率过大导致次优的优化结果，在前 5 个 epoch 中采用了 warmup 学习率保持训练的平稳性。

2.4.4 消融实验分析

（1）expansion rate 对性能的影响

本部分内容讨论 expansion rate 对整体识别模型性能的影响。根据本章 2.3.1 节所述，在设计 Broad Block 时，设置了一个以 1×1 卷积实现的输入扩张层，该层以一定的比例参数 expansion rate 对输入特征图进行冗余化，其作用是增加特征图的通道数，提高网络的宽度，以提升后续多分支结构的接受的特征丰富度。本节希望找到 expansion rate 与性能提升之间的关系。基于 YOLOX 框架，使用不同 expansion rate 下的 BroadNet 作为主干网络，在 Wildlife 测试集上的性能表现对比如表 2.3 所示。

表 2.3 不同 expansion rate 下 BroadNet 的性能表现对比

Model	Expansion rate	AP50 (%)	mAP (%)	Params (M)	FLOPs (G)
BroadNet	1.0	67.74	39.44	5.24	11.96
BroadNet	2.0	71.80	42.3	6.52	14.53
BroadNet	3.0	74.27	43.94	7.81	17.10

Expansion rate 的大小与性能的好坏在一定范围内呈现正相关。把 expansion rate 从 1 提高到 2，即将多分支结构的网络宽度增加一倍，AP50 从 67.74%提高到 71.8%，约 4%的提升，相应地 mAP 也提高了约 2.86%。性能的提升得益于网络参数量增加带来的更加丰富的特征。但网络宽度每增加一倍，便增加约 2.57G 浮点数计算量（FLOPs）以及相应的运行时空间代价。Expansion rate 的存在使得 BroadNet 成为一个可以缩放的网络。由于整个网络都由相同的 Broad Block 堆叠而来，Expansion rate 实际统筹了整个网络的参数量和宽度。出于对性能和运算代价的权衡考虑，在本文的对比实验中，选取 expansion rate=2.0 的对应模型与其他模型对比。

（2）多分支结构中不同路径对性能的影响

本章设计的特征提取模块 Broad Block 在训练时采用了多分支结构以及多个不同尺度的卷积核进行特征提取，本节对不同特征提取分支对性能的影响进行消融分析，目的是找到性能最优的多分支结构。

表 2.4 Broad Block 中多分支结构的不同路径对网络性能的影响

Structure	kernel size	AP50 (%)	mAP (%)
仅 7×7 conv	[7, 7, 7, 7]	69.98	40.79
+avg pooling		70.56	41.70
+asymmetric conv		71.15	42.16
+1×1 conv		71.8	**42.32**
+shortcut		**72.23**	41.97
+another 7×7 conv		71.66	42.06

由表 2.4 可见，对于 kernel size=7 的大卷积核网络，从只有单个 7×7 路径开始，依次增加非对称卷积（3×7、7×3），7×7 平均池化，以及 1×1 卷积都为模型带来了不同程度的增益。多分支结构整体带来了 2.28%的 AP 提升和 1.67%的 mAP 提升。在设计时，本文也考虑了 shortcut 连接的影响。在 Broad Block 中增加 shortcut 连接可以进一步提高网络的 AP50，但对 mAP 并没有帮助，反而可能有轻微的负面影响。因此，本文最终采用的多分支结构包括一个 7×7 路径、一个 3×7 路径、一个 7×3 路径，以及一个 7×7 平均池化和 1×1 路径。

（3）不同融合方法对多分支结构性能的影响

多分支结构的不同支路输出经过融合后送入下一模块，融合的方法会对网络模型的效果产生影响，本节对不同的融合方法的效果进行讨论，目的是找到一种最佳的融合方法。各种融合方法的性能对比如表 2.5 所示。

表 2.5 多分支结构的不同融合方法对比

融合方法	AP50 (%)	mAP (%)
直接相加融合	71.8	42.3
基于 SoftMax 归一化融合	72.1	42.45
基于 ReLU 归一化融合	71.29	42.2
无限制融合	70.42	41.23
基于绝对值归一化融合	**72.7**	**43.1**

以无参的直接相加方法作为基准，基于 SoftMax 归一化融合方法提高了约 0.15%的 mAP，性能超过直接相加的方法，但不太明显。而基于 ReLU 归一化融合方法、无限制融合方法与直接相加融合方法性能相当，甚至存在一定程度的性能下降，没有表现出明显优势。基于绝对值归一化融合方法展现出了较高的性能提升，mAP 高于直接相加约 0.8%。整体而言，含参的自适应融合方法对 BroadNet 的性能提升具有重要作用，基于绝对值归一化的融合方法表现最好。

2.4.5 对比实验分析

通过实验，本节进行两个层面的对比：一是将本章提出的 BroadNet 与其他主干网络进行横向对比，包括几个高性能主干网络以及轻量化主干网络；二是将本章基于 BroadNet 改进的整体野生动物识别方法与其他识别方法进行纵向对比，衡量本章方法的整体改进效果。在进行对比实验时，考虑到网络在不同分辨率下的泛化性，在 Wildlife 野生动物测试集上，本节分别在 416×416 以及 608×608 两个分辨率上进行了推理。为了减小随机性，保证对比的效果，在训练时，未使用多尺度训练。本节用作与其他网络对比的 BroadNet 的 expansion rate=2.0，kernel size=7，每个阶段包含一个 Broad Block 特征提取模块，参数层共有 20 层。

1. 与其他主干网络的对比

为了对 BroadNet 的性能进行评估，本小节将 BroadNet 与其他主干网络的推理性能进行了对比，包括 YOLOX 使用的主干网络 CSPDarkNet，DarkNet，以及 ResNet 等高性能网络，以及一些轻量级主干网络。在对比时，基于 YOLOX 识别框架，仅做主干网络的替换，保持变量唯一。本节还进一步通过可视化的方式对比了 BroadNet 与其他主干网络的有效感受野。

（1）与高性能主干网络的性能对比

CSPDarkNet、DarkNet 等是 YOLO 系列识别模型常用的特征提取主干网络，有必要与这些高性能网络进行比较，观察性能和效率的优劣。BroadNet 与 CSPDarkNet、DarkNet 等高性能网络对比如表 2.6 所示。

表 2.6 BroadNet 与高性能主干网络的推理性能对比

主干网络	输入分辨率	AP50(%)	mAP(%)	Params(M)	FLOPs(G)
BroadNet	416×416	**71.80**	**42.32**	**6.52**	**14.53**
CSPDarkNet[30]	416×416	68.58	41.58	18.55	29.28
DarkNet-21[24]	416×416	65.98	39.67	26.99	24.92
ResNet-34[32]	416×416	63.73	37.51	21.8	25.32
BroadNet	608×608	**63.7**	**29.4**	**6.52**	**31.03**
CSPDarkNet	608×608	55.9	27.0	18.55	62.37
DarkNet-21	608×608	43.9	20.9	26.99	53.24
ResNet-34	608×608	49.1	22.2	21.8	54.09

从表 2.6 可以看出，BroadNet 具有最高的识别性能，超过 CSPDarkNet、DarkNet 等目标识别任务中的先进主干网络，也超过 ResNet-34 等经典主干网络。值得注意的是，此处对比使用的网络深度和参数量都高于 BroadNet。在分辨率为 416×416 时，BroadNet 的 mAP 比 CSPDarkNet 高 0.88%，AP50 高约 3.22%，参数量是 CSPDarkNet 的三分之一。同时，FLOPs 也低于 CSPDarkNet 和 DarkNet-21。当分辨率提高到 608×608 时，BroadNet 的 mAP 比 CSPDarkNet 高 2.4%，AP50 则高约 7.8%，FLOPs 约为后者的一半，差距更为明显。大卷积核的设计带来了更大的输入分辨率鲁棒性。

（2）与轻量级主干网络的性能对比

在表 2.7 中，本文将 BroadNet 与 MobileNet-V2[63]，ShuffleNet-V2[64]，ResNeXt-34 等轻量级主干网络进行了对比。由表可知，轻量级主干网络虽然参数量更少，计算量也更低，但是性能差于 BroadNet。在分辨率为 416×416 时，ResNeXt-34 与 BroadNet 的性能非常接近，但当分辨率增大到 608×608 时，ResNeXt-34 与 Broadnet 的性能差距进一步增大了。其他轻量级主干网络如 MobileNet-v2、ShuffleNet-v2 等识别效果与

BroadNet 差距较大。BroadNet 比轻量级主干网络性能更强，但参数量和计算量等并未高太多。

表 2.7 BroadNet 与轻量级主干网络的推理性能对比

主干网络	输入分辨率	AP50(%)	mAP(%)	Params(M)	FLOPs(G)
BroadNet	416×416	**71.80**	**42.32**	6.52	14.53
MobileNet-v2[63]	416×416	68.90	39.0	4.45	7.39
ResNeXt-34[34]	416×416	70.03	40.62	1.52	2.61
ShuffleNet-v2[64]	416×416	56.73	30.11	**0.94**	**1.05**
BroadNet	608×608	**63.7**	**29.4**	6.52	31.03
MobileNet-v2	608×608	54.90	23.36	4.45	15.79
ResNeXt-34	608×608	61.0	27.2	1.52	5.57
ShuffleNet-v2	608×608	45.50	18.50	**0.94**	**2.32**

（3）与其他主干网络的有效感受野对比

本节将 BroadNet 与其他主干网络的有效感受野进行定性分析，目的是对感受野有一个直观的对比。根据文献[41]给出的有效感受野计算方式，本文将 ShuffleNet、CSPDarkNet、ResNet-34 以及 BroadNet 的有效感受野进行可视化，如图 2.16 所示。深色部分颜色越深，范围越大，表明有效感受野越大。由于深度和卷积核大小都会对有效感受野产生影响，图 2.16 给出了 Kernel size∈{5，7，11}设置下，深度为 20 层的 BroadNet 的感受野可视化结果。

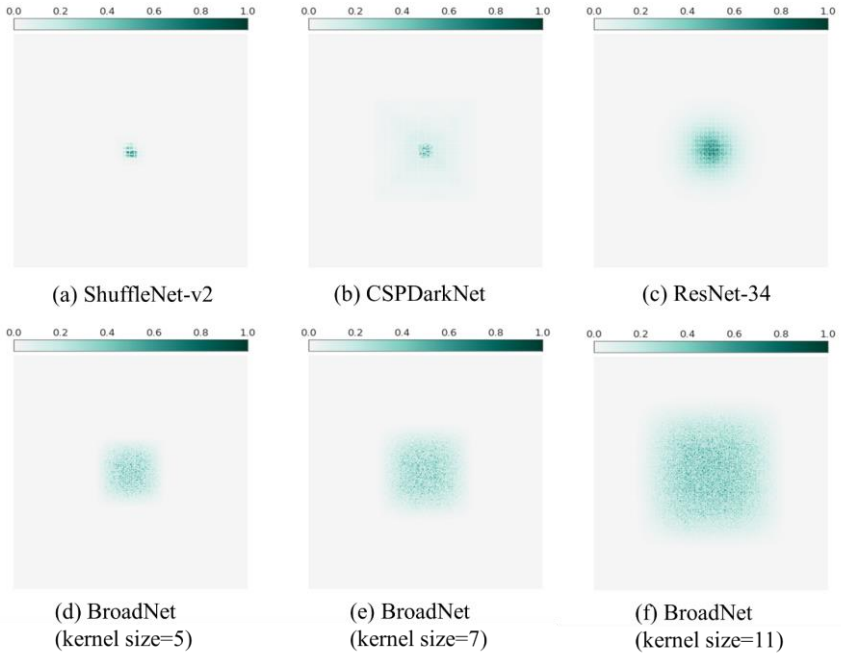


图 2.16 不同网络的有效感受野对比，深色部分越大，有效感受野越大

从图 2.16(a)、图 2.16（b）、图 2.16（c）可以看出，ShuffleNet-v2 具有较小的有效感受野，而 CSPDarkNet 的有效感受野整体范围较大，但响应区域不均匀，只有中间的小部分区域响应强烈。ResNet-34 的感受野相对比较均匀，整体范围虽然比 ShuffleNet 大，但小于 CSPDarkNet。相比之下，BroadNet 在不同卷积核尺度下都具有大且更均匀的有效感受野，这意味着 BroadNet 可以在更大范围内充分关注各空间位置的图像特征。从图 2.16(d)到图 2.16（e）再到图 2.16（f），增大卷积核的尺度直接带来了有效感受野的提升。当 kernel size=7 时，BroadNet 的有效感受野已经大于其他网络，进一步增大到 11，感受野的差距则更为明显。

2. 与其他野生动物识别方法的对比

本节将本章方法与其他野生动物识别方法的识别性能进行对比，目的是验证本章方法在野生动物场景下的适用性。表 2.8 给出了 YOLOX、YOLOv4、YOLOv5 等方法在 416×416 分辨率下的推理性能。为了对比的公平性，本文尽可能选取了识别效果、参数量和计算量比较接近的实现。

如表 2.8 所示，在 Wildlife 测试集上，本章方法超过了基准模型 YOLOX、YOLOv4、YOLOv5 等其他几种较新的 YOLO 系列方法，在性能和参数量上都具有优势。FCOS 这类高性能目标识别网络虽然识别性能很好，但计算量太大，本章方法与之相比，具有更少的推理时参数量和浮点数计算量。在野生动物场景下，本章方法具有一定的适用性。

表 2.8 本章方法与其他识别方法的推理性能对比

Method	AP50 (%)	mAP (%)	Params (M)	FLOPs (G)
本章方法	**71.80**	**42.32**	**17.54**	39.39
YOLOX	68.58	41.58	29.57	54.14
YOLOv4[25]	67.82	40.22	41.19	29.56
YOLOv5[27]	68.27	40.64	21.09	**21.42**
FCOS[22]	71.21	42.27	32.13	68.36

2.4.6 识别结果展示及问题分析

（1）实验结果展示

本节给出本章野生动物识别方法的一些识别结果，如图 2.17 所示。在原始 YOLOX 下，存在一些样本无法正确检出，红框内标出了漏检的目标。最左边的大象样本存在目标不全的问题，原始 YOLOX 无法识别出此类目标，右上角大象目标会被漏检，而本章方法可以给出精确的识别框。中间的马群样本和右边的老虎样本存在部分遮挡问题，YOLOX 会漏检，而本章方法可以精确识别。

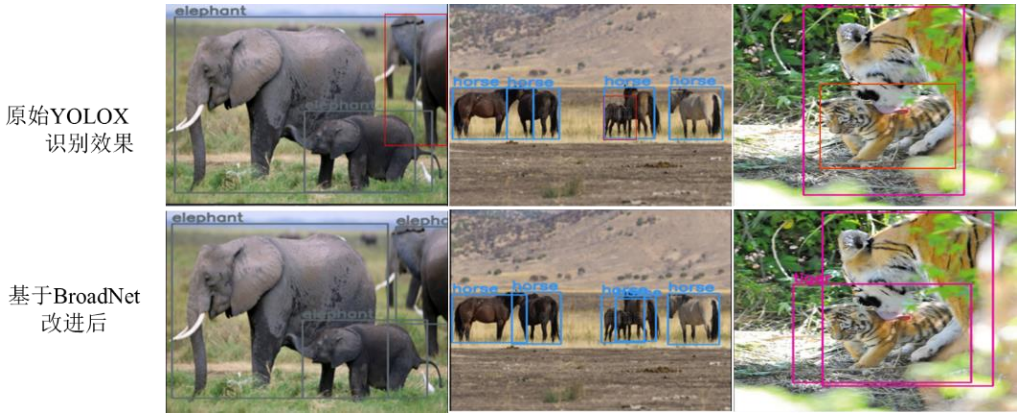


图 2.17 部分样本识别结果对比

（2）存在的问题

本章方法虽然提高了整体的野生动物识别效果，在一定程度上好于原始 YOLOX，但依然会出现漏检以及定位不准的情况，如图 2.18 所示。在第三章，本文通过构建空间注意力模块和通道注意力模块来对有用的目标区域进行关注，提出一种基于注意力机制的野生动物识别方法，提高定位和识别精度。

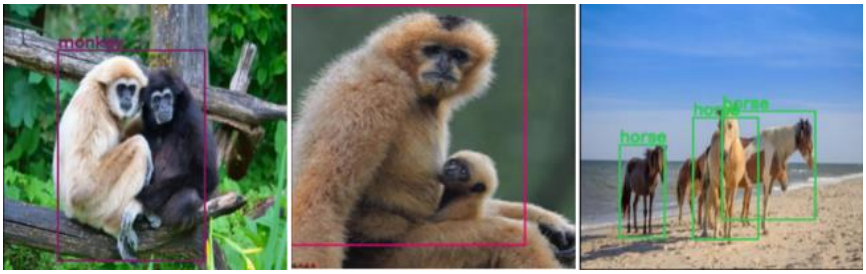


图 2.18 本章方法改进后依然存在问题的样本示例

2.5 小结

本章选取了在公开数据集上表现优秀的通用识别网络 YOLOX 作为野生动物识别任务的基准模型。然而，YOLOX 的识别效果依然不够理想。为了解决 YOLOX 在部分野生动物场景下无法有效识别目标、漏检等问题，本章提出了一种基于结构重参数主干网络的野生动物识别方法。该方法设计了主干网络 BroadNet，通过 7×7 的大卷积核来增大神经网络的感受野，并通过深度卷积块来减小卷积核增大带来的参数量和计算量。在训练时，使用多分支结构提高网络的表达能力，推理时转化为简洁的直筒式结构，在参数量较少、计算量较小的情况下提高了网络的感受野。本章提出的方法可以有效识别一些 YOLOX 漏检、识别不准的样本。在自建野生动物数据集 Wildlife 上，本章方法具有较高的识别准确率，超过了 YOLOv4、YOLOv5 等其他高效通用目标识别模型，但依然有一些样本无法有效识别和定位，存在进一步的改进空间。

第三章 基于注意力机制的野生动物识别方法研究

3.1 概述

在第二章，本文介绍了一种基于结构重参数主干网络 BroadNet 的野生动物识别方法，解决了通用识别网络 YOLOX 部分野生动物样本在遮挡、目标不全情况下漏检的问题，但该方法依然存在不足。由于野生动物场景存在背景丰富、纹理信息复杂等特点，第二章方法对部分困难样本无法精确定位和识别，仍存在漏检、定位不准等问题，有进一步改进的空间。本章认为在野生动物识别模型中引入注意力机制（Attention Mechanism）可以在一定程度上解决这个问题，让网络更加关注前景目标区域，提高对困难目标的识别效果。

注意力机制是一类让模型自适应地关注重要特征区域的方法。对于野生动物识别这种具有复杂背景和多变姿态的场景，注意力机制具有很大的应用空间。Wang 等[45]提出了一种非局部网络（Non-local Neural Networks）来捕捉全局空间相关关系，在输入特征图上计算不同位置之间的相似度，通过归一化得到注意力权重，动态关注不同的空间位置。CCNet[46]通过新颖的 Criss-cross attention 模块获取交叉路径上其周围像素的上下文信息，提出了一种可以更加高效地捕捉长距离依赖信息的方式。SENet[47]使用了如全局平均池化和通道降维等技术，关注特征图中的重要通道。ECANet[48]通过一维卷积层来关注通道维度上的局部相关性，生成注意力掩码来关注重要通道。CBAM[49]联合使用空间注意力和通道注意力机制，通过池化等手段聚合空间信息，在公开数据集上取得了良好的表现。但是它的全局最大池化和全局平均池化会损失空间不同位置的语义相关性，因此在前景目标分布多样的野生动物识别任务中可能表现不佳。本文认为，可以对 CBAM 进行改造，设计一种适应野生动物场景特点的注意力机制模块作为提高性能的手段。

为了解决在复杂场景下部分野生动物样本定位不准、漏检等问题，基于第二章工作，本章提出一种基于注意力机制的野生动物识别方法。该方法结合了双向特征融合网络 PAFPN，应用了本章设计的联合注意力模块 Ultra-CBAM，通过实验选取了最优注意力插入位置，进一步提高了在野生动物场景下的识别效果。Ultra-CBAM 综合了 CBAM 的优势并加以优化，基于一维卷积子网络，提高了通道注意力跨通道的信息交互能力，使用参数化的方式优化了空间注意力聚合通道信息的过程。本章提出的野生动物识别模型在 Wildlife 测试集上 mAP 达到 48.90%，AP50 达到 76.51%，进一步提高了第二章方法的性能，对一些第二章方法漏检、识别不准的样本具有良好的识别效果。

3.2 相关工作

3.2.1 注意力机制

注意力机制的基本思路是通过学习特征图不同的区域之间的相关性来生成掩码，增强重要区域。注意力机制包括多种，本文关注两个维度的注意力机制，一是通道维度，二是空间维度。

深层特征图中的不同通道可能代表了不同的对象，对不同通道进行不同权重的注意力分配，就可能实现对更加重要对象的关注。SENet[47]是通道注意力机制的代表。SENet 关注特征图中一直被忽视的通道维度，它使用全连接层和全局平均池化操作生成逐通道的注意力权重掩码向量。SENet 的核心步骤包括信息聚合和注意力生成。信息聚合的过程通过全局平均池化的方法来收集每个通道的特征图的全局统计信息，而注意力生成过程则对通道间的相关性进行捕捉，生成一个逐通道的注意力掩码向量。将生成的注意力掩码与原始特征图进行相乘，得到对每个通道进行不同比例缩放后的特征图。SENet 以一定比例对输入的通道注意力向量进行降维操作，嵌入到一个低维空间中，再进行反向变换，在变换的过程中完成对通道间信息的关注。

Park 等[50]提出了 BAM，使用并行的通道注意力和空间注意力分支同时从特征图中生成空间掩码和通道掩码，利用空洞卷积扩大空间注意力子模块的感受野。基于 BAM，CBAM[49]以一种更高效的方式综合了通道注意力和空间注意力机制，通过依次提取特征图的空间注意力和通道注意力实现对空间维度和通道维度的高效关注。CBAM 的通道注意力模块基于 SENet，如图 3.1 所示，通过全局平均池化和全局最大池化聚合空间信息，并使用共享的全连接网络进行降维，生成通道注意力向量，基于学习到的通道注意力权重对不同通道的特征图进行加权。

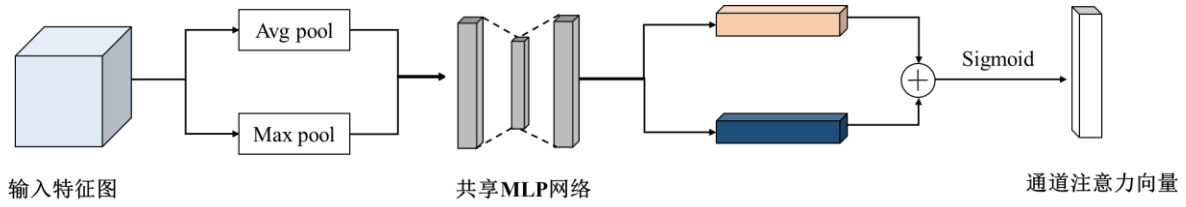


图 3.1 CBAM 中的通道注意力机制

CBAM 的空间注意力模块通过全局平均池化和最大池化聚合通道信息，并基于一个卷积层来生成空间注意力图，为不同的空间位置赋予不同的权重，如图 3.2 所示。

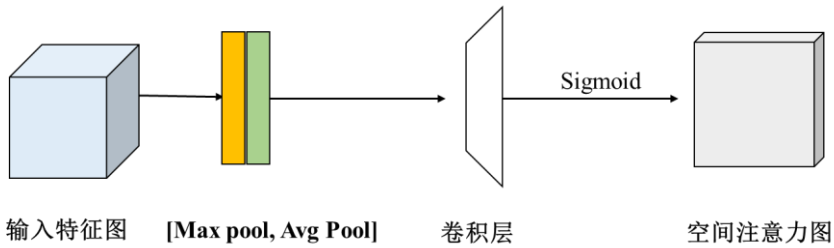


图 3.2 CBAM 中的空间注意力机制

CBAM 是一种优秀的注意力机制，它结合了对空间维度和对通道维度的注意力模块，通过分别生成注意力掩码来突出重要空间区域和通道。在公开数据集上，CBAM 具有良好的表现，因此，本章考虑将其引入野生动物场景中，目的是提高对特征图中前景目标的响应。但是在接下来的实验中，该方法在野生动物数据集 Wildlife 上的效果不够理想，存在提升空间。

3.2.2 注意力机制 CBAM 在野生动物场景下的表现

本节基于实验，观察将 CBAM 引入野生动物场景下的表现，如表 3.1 所示。在其他实验设置保持相同的情况下，将 CBAM 放入特征融合网络中增强不同尺度特征图的融合过程。本节给出了两种不同卷积核大小下 CBAM 的效果，CBAM-3-8 表示空间注意力卷积核大小为 3、通道注意力降维比例为 8 的 CBAM 模块，CBAM-7-8 则使用了大小为 7 的卷积核来生成空间注意力。可以看出，引入 CBAM 可以带来约 0.5%左右的 AP 增益，而 mAP 效果则不明显。这表明 CBAM 无法直接迁移到野生动物场景下，本文就此进行针对性的分析和改进。

表 3.1 CBAM 在野生动物场景下的表现

Attention Module	AP50 (%)	mAP (%)	Params (M)	FLOPs (G)
无注意力	75.65	48.27	**26.15**	**46.28**
CBAM-3-8	**76.21**	48.21	26.18	46.28
CBAM-7-8	75.73	**48.33**	26.23	46.28

就空间注意力而言，CBAM 中使用了全局最大池化等操作聚合通道信息，这可能使得不同的空间位置处得到的响应来自于不同的通道，破坏了每个通道上空间局部特征之间的关联，也不能在训练过程中自适应地调整信息聚合过程。同时，生成空间注意力图的过程仅使用了一个简单卷积层，但野生动物场景复杂，背景多样，导致这种方式对有用区域的关注不够充分，存在改进空间。

就通道注意力而言，CBAM 沿用了 SENet 使用共享全连接网络生成注意力向量的思路，同时也具有 SENet 的问题。ECANet[48]认为避免使用全连接网络降维有助于学习更有效的通道注意力，适当的跨通道交互可以在显著降低模型复杂度的同时保持性能。该方法使用单层的一维卷积代替了全连接层，关注通道维度上局部窗口内不同通道的关联，在公开数据集上取得了超过 CBAM 的性能，本文认为 ECANet 依旧存在不足，可以进一步改进。

3.3.3 堆叠式 ECANet 设计

ECANet 通过单层一维卷积来捕获局部跨通道的关联信息。单层的一维卷积，受限于其卷积大小，即便是根据输入通道数自适应确定窗口，也可能无法捕获范围足够大的通道间信息，限制了网络的性能。因此，本文提出使用多层堆叠的一维卷积子网络在更大范围内捕捉跨通道的信息相关性。

本文通过一个简单的实验进行了验证，构建一个一维卷积核大小为 11 的单层 ECANet，以及一个具有 3 个大小为 7 的卷积核的堆叠 ECANet，对比两者的效果。如表 3.2，在其他实验设置相同的前提下，多层堆叠的方式在几乎没有额外参数量和计算量的情况下，取得了约 3%AP50 提升以及约 1.7%mAP 提升。这表明使用堆叠的一维卷积子网络可以提高通道注意力的性能。

表 3.2 单层 ECANet 与具有多个堆叠卷积层的 ECANet 对比

Method	AP50（%）	mAP（%）
Raw ECA-11	72.25	46.59
Stack ECA-7	**75.24**	**48.31**

一维卷积核的大小 k 与通道数 C 成正相关[48]，由于通道数通常为 2 的幂次，ECANet 引入指数函数，构建了一个 k 和 C 之间可能的映射关系：

$$k=\psi(C)=\left|\frac{log_2(C)}{\gamma}+\frac{b}{\gamma}\right|_{odd} \tag{3.1}$$

其中 γ 、b 为超参数。

3.3 基于注意力机制的野生动物识别方法

针对第二章方法对部分样本识别时存在的定位不准、漏检等问题，本节提出一种基于注意力机制的野生动物识别方法。该方法使用本章设计的新颖通道-空间联合注意力机制 Ultra-CBAM 对特征融合过程进行增强。Ultra-CBAM 基于 CBAM，进行了适应野生动物场景的改进，突出前景目标，增强前景像素的特征图响应。接下来对整体识别方法进行概述，再依次介绍各个子模块的设计。

3.3.1 整体设计

本章在第二章方法的基础上进行改进。如图 3.3 所示，本章方法由 BroadNet 主干网络、特征融合网络、预测头几大部分组成。第二章使用了 FPN 作为特征融合网络对 C3、C4、C5 三个层级的特征图进行融合，这种方式只有自顶向下的融合路径，无法充分利用浅层特征图。因此，本章引入了具有自底向上融合路径的 PAFPN[31]提升特征融合效果。并进一步利用本章设计的通道-空间联合注意力机制 Ultra-CBAM 对特征融合过程进行增强，得到一种基于注意力机制的特征融合网络 ATT-PAFPN。

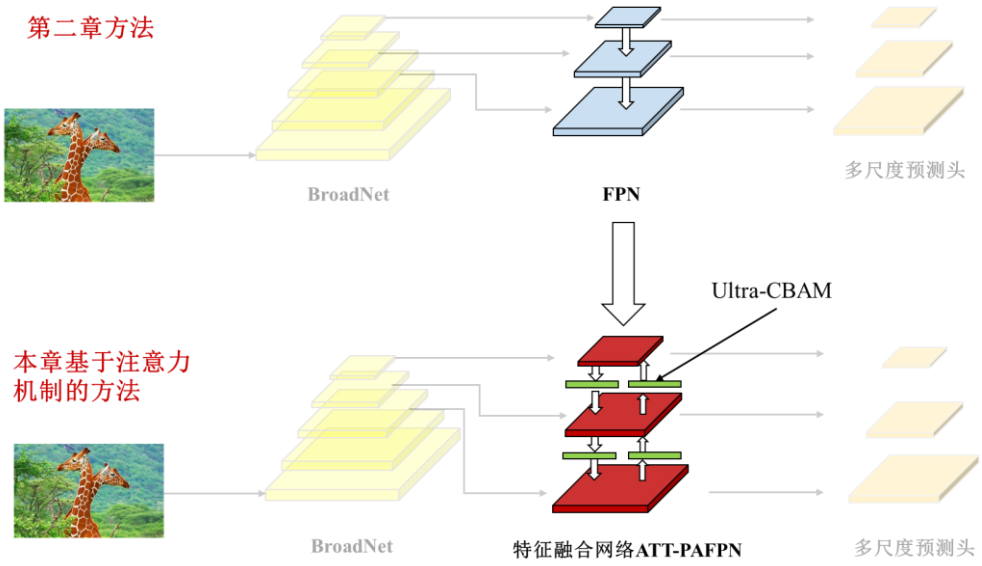


图 3.3 基于注意力机制的野生动物识别方法

ATT-PAFPN 可以在融合不同层级的特征图过程中提取特征图的通道和空间相关性并为重要区域分配更大的比重，为后续对目标类别和位置的预测准备更高质量的特征图。ATT-PAFPN 基于 Ultra-CBAM，它综合了 CBAM 的优势并加以优化使其更适应野生动物场景。基于分层的一维卷积子网络，提高通道注意力跨通道的信息交互能力。同时使用参数化的方式优化了空间注意力聚合通道信息的过程，减小了聚合过程中的信息损失，基于分组卷积核和 1×1 卷积，优化了生成空间注意力图的过程。接下来对 ATT-FPN 的具体设计进行阐述。

3.3.2 特征融合网络 ATT-PAFPN 设计

如图 3.4 所示，ATT-PAFPN 的输入为由 BroadNet 主干网络提取的 C3、C4、C5 三个尺度不同、语义信息不同的特征图，输出经过自上而下和自底向上两个路径融合后的三个特征图。经过两个路径的特征融合以及融合过程中注意力的监督，特征融合网络 ATT-PAFPN 的输出特征图既突出了前景目标区域，又结合了不同语义层级的信息。ATT-PAFPN 以 PAFPN 为基础，结合本章设计的联合注意力机制 Ultra-CBAM 对特征融合过程进行增强，可以输出具有更强目标区域响应的特征图，便于后续预测头进行识别和定位。

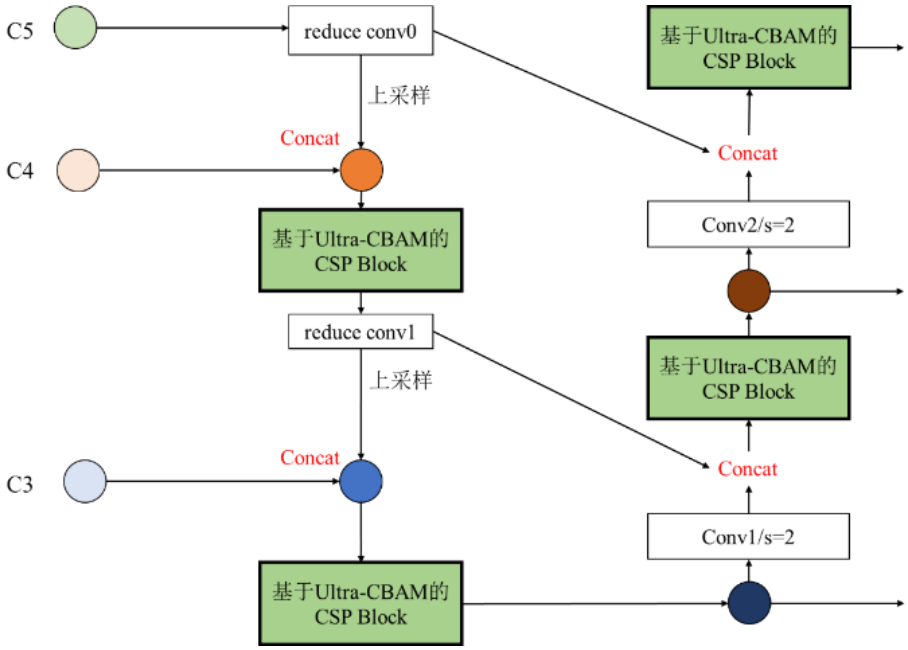


图 3.4 注意力增强的特征融合网络 ATT-PAFPN

（1）双向特征融合网络 PAFPN

第二章方法使用了 FPN[28]作为特征融合网络。同样是以 C3、C4、C5 三个层级的特征图作为输入，FPN 只有自顶向下的融合路径，仅仅对浅层特征图进行增强，深层特征图表达能力不强。而 PAFPN 在 FPN 的基础上再次进行了自底向上的融合，通过跨层连接，将高层特征和低层特征进行融合，增加了特征图语义信息的丰富度。在自上而下的路径中，将 C5 特征图通过横向连接卷积层 reduce conv0 进行通道降维，这是因为 C5 特征图包括了大量冗余通道，对其进行压缩可以减小融合网络的计算量。降维后通过上采样得到与 C4 特征图宽高相同的特征图，将两者在通道维度上进行拼接，就将两个不同尺度的特征图进行了融合。其余层级的特征图融合同理。自底向上的融合路径使用一个步长为 2 的卷积层来对注意力增强后的 C3 特征图下采样，下采样后的 C3 特征图也通过通道维度拼接的方式与 C4 特征图进行融合。

（2）PAFPN 中的注意力插入方案设计

在 PAFPN 的特征融合过程中，存在多个不同位置可以使用注意力模块进行增强，但不同位置下取得的性能提升存在差异。对应图 3.5 中的不同位置，在设计时，本文考虑了以下几种方案：

A. 插入 PAFPN 的输入阶段。直觉上，由于输入特征为从原始特征图的不同阶段所得的，特征融合模块对于输入特征图非常敏感，在此处插入注意力机制可能会影响特征融合的效果。

B. 插入 CSP Block 中。CSP Block 是 PAFPN 中的主要特征提取模块，在 CSP 模块中应用注意力机制可以同时利用注意力特征图以及原始特征图，得到信息更丰富的特征图输出。

C. 插入每级特征拼接后。不同阶段的特征图经过上/下采样后，与其他阶段的特征

图进行拼接。将拼接后的特征图送入注意力模块，可能也有助于提升特征融合效果。

D. 插入 PAFPN 的输出阶段。

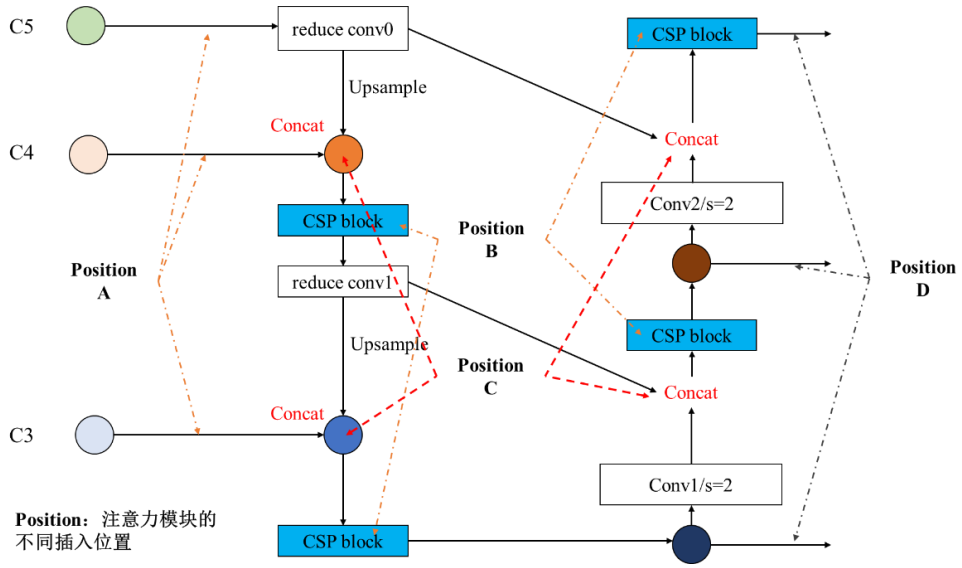


图 3.5 PAFPN 中注意力模块增强位置的不同设计方案

根据本章 3.4 节消融实验，将 Ultra-CBAM 放在 CSP Block 中性能最佳，因此本章最终采取了这种方案。基于 Ultra-CBAM 的 CSP Block 如图 3.6（a）所示，在 CSP Block 中，输入特征图有两个流向：一部分经过卷积后被降维，送入若干 BottleNeck 块提取特征。另外一部分也进行降维，并直接与前者的输出进行拼接。这样降低了 BottleNeck 块输入的特征图通道数量，既可以减少参数量，又优化了梯度流向，使得网络更易于学习。Ultra-CBAM 对 BottleNeck 块的输出进行通道注意力和空间注意力的提取和关注。将 Ultra-CBAM 增强后的特征图与另一支路进行拼接融合，同时利用注意力特征图以及原始特征图可以得到信息更丰富的输出特征图。接下来本文对 Ultra-CBAM 的具体设计以及其两个子模块结构进行阐述。

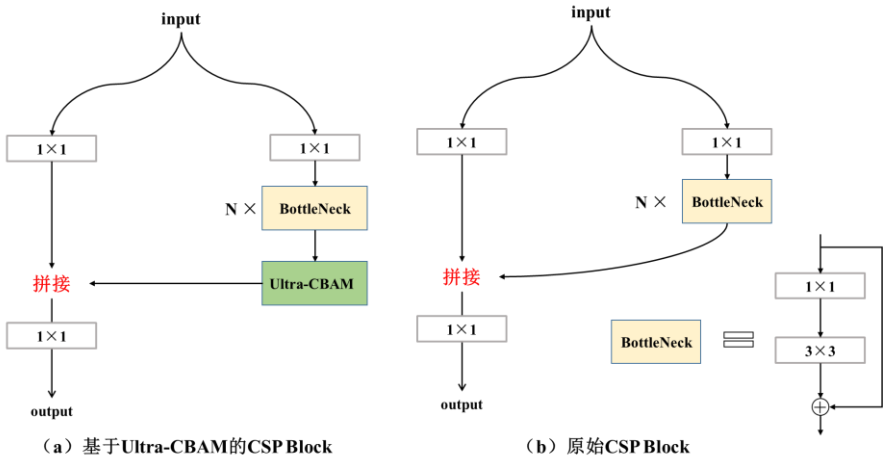


图 3.6 基于 Ultra-CBAM 的 CSP Block 结构

3.3.3 通道-空间联合注意力模块 Ultra-CBAM 设计

通道-空间联合注意力模块 Ultra-CBAM 整体结构如图 3.7 所示。Ultra-CBAM 基于

CBAM[49]改进，延续了其基本结构，将空间注意力机制和通道注意力机制分离，按照串联的顺序依次对输入特征图进行通道注意力与空间注意力的提取及关注。Ultra-CBAM 对于前景目标形态复杂、背景信息杂乱的野生动物识别场景具有更好的适用性。接下来分别对通道注意力子模块 Cascade-CAM 和空间注意力子模块 Ultra-SAM 的设计进行阐述。

Ultra-CBAM 包括对通道注意力子模块的改进设计 Cascade-CAM（Cascade Channel Attention Module）以及对空间注意力子模块的改进设计 Ultra-SAM（Ultra Spatial Attention Module）。Cascade-CAM 基于一维卷积分层子网络，对不同通道的局部相关性进行关注。Ultra-SAM 基于 1×1 卷积进行通道信息的聚合，最大程度地减少了聚合过程中的语义信息损失，使用分组卷积、平均池化和最大池化来生成空间注意力图。相比于 CBAM，Ultra-CBAM 减少了在聚合信息生成注意力过程中的信息损失，更加关注特征图的不同通道之间的局部关联性，可以更好地突出前景目标。

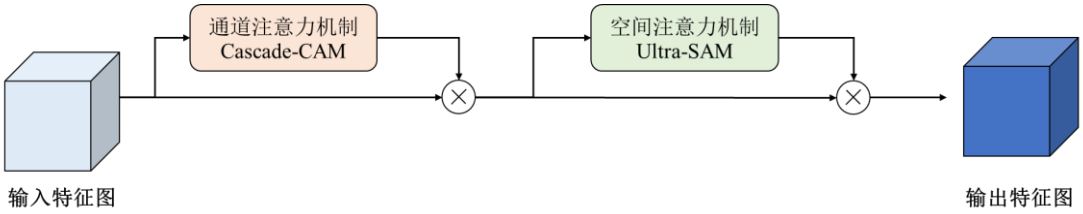


图 3.7 Ultra-CBAM 模块整体结构

（1）通道注意力子模块 Cascade-CAM 设计

一个深层特征图的不同通道代表了不同的对象。网络越深，特征图的通道数越多，必然存在一些通道是冗余的。为含有区分性目标区域的通道赋予更大的权重，降低冗余通道的比重，就可能让网络更加关注有用信息，带来性能的提升。由本章 3.3.3 节所述，堆叠式一维卷积子网络可以为通道注意力 ECANet 带来较大的性能提升。基于此先验知识，本节将堆叠式一维卷积子网络设计进一步引入 Cascade-CAM 中。Cascade-CAM 是一种基于堆叠式一维卷积子网络的通道注意力模块。它可以为特征图通道重新分配比重，提高特征图的表达能力，Cascade-CAM 结构如图 3.8 所示。

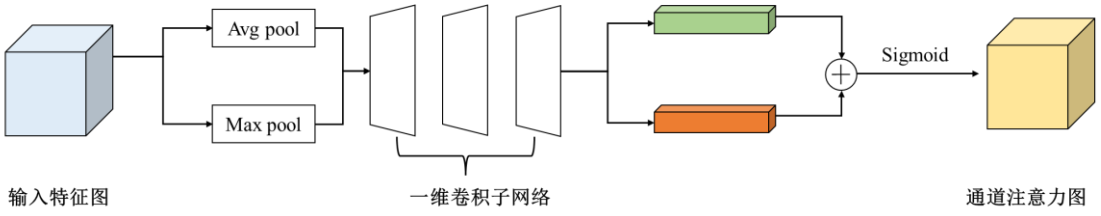


图 3.8 通道注意力模块 Cascade-CAM 结构设计

给定输入特征图 F ,首先通过全局平均池化和全局最大池化操作对空间信息进行聚合，得到两个不同的空间上下文描述子 F_{avg}^{c} 和 F_{max}^{c} ，分别代表全局平均池化得到的中间特征和全局最大池化得到的中间特征。全局平均池化保留了特征图的空间上不同位置的信息，而全局最大池化则保留了每个通道上的最大响应，两者互补。将这两个描述

子分别送入参数共享的一维卷积子网络，产生了各自的通道注意力掩码。将注意力掩码逐元素相加融合后，使用 Sigmoid 函数进行激活，就得到了通道注意力向量 $M_c \in R^{C\times1\times1}$。通道注意力向量会为那些不具有区分性的通道赋予更小的权重，增强更重要的通道。Cascade-CAM 生成通道注意力向量的过程可以表示为

$$M_c(F)=\sigma\left(Subnet\left(AvgPool(F)\right)+Subnet\left(MaxPool(F)\right)\right) \tag{3.2}$$

其中，*Subnet* 表示一维卷积子网络，σ 表示 Sigmoid 激活函数。相比之下，CBAM 中的通道注意力模块采用了全局平均池化和全局最大池化来进行空间信息的聚合，使用了全连接网络来生成通道注意力向量，按照一定降维比例，将通道注意力向量嵌入到低维空间中以减少参数量，这种方式效率较低。本文使用多个一维卷积层堆叠形成级联的子网络，扩大了通道注意力窗口，提高了通道间信息交互的能力。使用一维卷积子网络生成注意力向量的方式可以更好地捕捉通道局部相关性。

一维卷积子网络的卷积核大小可由式 3.1[48]确定。网络可以自适应地根据输入特征图的通道数来选择局部跨通道交互的窗口大小 k。在 3.4 节，本文将进一步对影响 Cascade-CAM 的因素进行讨论。接下来对 Ultra-CBAM 中的空间注意力机制 Ultra-SAM 进行介绍。

（2）空间注意力子模块 Ultra-SAM 设计

CBAM 中利用了全局平均池化和全局最大池化来聚合通道信息。这种方式计算简单快速，但是可能损失信息。Ultra-SAM 对 SAM 进行了进一步的增强，对通道信息聚合的过程以及生成空间注意力的过程都进行了改造和创新，使用一个具有 K×K 卷积核的分组卷积、平均池化和最大池化来对空间位置上的信息进行关注。其过程如图 3.9 所示。

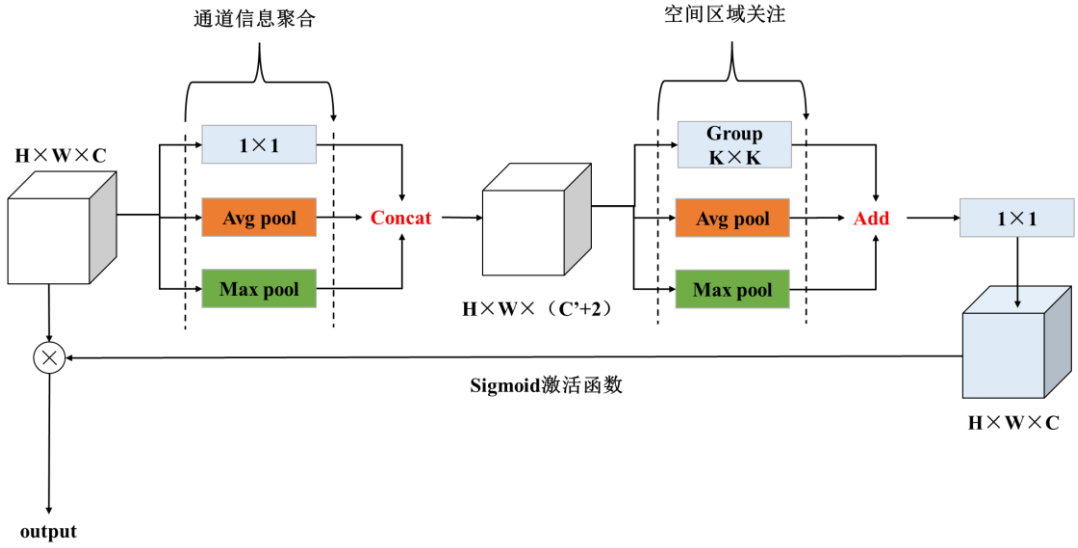


图 3.9 Ultra-SAM 结构设计

同类工作 CBAM 中空间注意力的通道信息聚合过程仅采用了全局平均池化和全局

最大池化。全局平均池化对于所有通道一视同仁，同一通道上的不同空间位置具有相同权重。而全局最大池化使得不同的空间位置处得到的响应来自于不同的通道。这可能会破坏空间上的局部特征之间的关联，除此之外，这两种池化都是无参的，不能在训练过程中自适应地调整信息聚合过程。因此，本文希望能有一种方式可以让网络对所有通道既能一视同仁又可以让网络可以自适应地进行通道间的信息聚合。

1×1 卷积是解决这个问题合适的工具。Ultra-SAM 在 CBAM 使用的全局平均池化和全局最大池化的基础上，增加一个 1×1 卷积的并行支路。它的作用是让网络自适应地关注通道间的关系，同时保留空间上各位置之间的关联。每一个支路的输出特征图通过拼接的方式进行聚合。本文设置了降维比例（reduction rate）r 来控制 1×1 卷积的输出通道数 C'，有 $C'=C/r$。拼接之后的特征图进入一个卷积块进行特征提取。用 F' 表示将输入特征图聚合通道信息后得到的中间特征图。F 表示原始的输入特征图，聚合通道信息的过程可以表示为

$$F'=\left[f^{1\times1}(F);MaxPool_{channel}(F);AvgPool_{channel}(F)\right] \tag{3.3}$$

其中，$f^{1\times1}$ 代表 1×1 卷积核。$MaxPool_{channel}$、$AvgPool_{channel}$ 分别代表在通道维度上的全局最大池化和全局平均池化。

Ultra-SAM 使用一个具有 K×K 卷积核的分组卷积、平均池化和最大池化来对空间位置上的信息进行关注。对他们的输出进行融合，得到具有更加丰富位置信息的空间特征。使用 1×1 卷积进行升维以及通道间信息聚合。将输出特征使用 sigmoid 激活，就得到空间注意力图 $M_s\in R^{1\times H\times W}$。Ultra-SAM 整体流程可以表示为

$$M_s(F)=\sigma\left(f^{1\times1}\left(sum\left(AvgPool_{space}(F'),MaxPool_{space}(F'),f_{Group}^{k\times k}(F')\right)\right)\right) \tag{3.4}$$

其中，$f^{K\times K}$ 代表 K×K 卷积核，K 为卷积核大小，取 3 或 7。σ 表示 sigmoid 函数。

3.4 实验与分析

本节通过实验验证本章方法在野生动物场景下的有效性并进行分析。主要有以下几个方面的内容：

1. 对本章提出的 Ultra-CBAM 的不同模块进行消融实验，确定最优的结构和参数，分析各个模块如何发挥作用。

2. 以第二章方法为基准模型，对比本章方法在野生动物场景下的有效性。

3. 对比本章设计的 Ultra-CBAM 方法和其他注意力方法，验证本章方法在野生动物场景下的适用性。

本章使用与第二章相同的评价指标 mAP 和 AP50 来对模型性能进行评估。数据集、数据增强方案、Batch size、学习率设置等也与第二章相同。接下来对消融实验和对比实验进行介绍。

3.4.1 消融实验分析

本节对 Ultra-CBAM 进行消融实验，确定最优的注意力结构和参数。先讨论对最优注意力增强位置的选择，再对影响 Cascade-CAM 的几个因素进行分析，最后讨论 Ultra-SAM 的影响因素。

（1） PAFPN 中最优注意力插入方案选择

本节给出在其他设置相同的情况下，让 Ultra-CBAM 模块在 PAFPN 的不同位置进行注意力监督所得到的实验效果，目的是找到一种效果最优的注意力插入方案。

表 3.3 将 Ultra-CBAM 插入 PAFPN 的不同位置时的性能对比

插入位置	AP50 (%)	mAP (%)
PAFPN 输入阶段	75.10	48.26
CSP Block	**76.26**	**48.62**
特征拼接后	73.76	47.93
PAFPN 输出阶段	75.44	47.94

如表 3.3 所示，将 Ultra-CBAM 置于 CSP Block 中取得了最高的 mAP 和 AP50。放在特征拼接后效果最差，因此本文最终选择将 Ultra-CBAM 置于 CSP Block 中进行注意力监督，基于 Ultra-CBAM 提高 CSP Block 的特征提取能力，进而提高整体特征融合网络的融合效果。

（2）Cascade-CAM 和 Ultra-SAM 消融分析

Ultra-CBAM 基于 CBAM，对其通道注意力子模块和空间注意力子模块都进行了改进，本节对这两个模块进行消融分析。基于 YOLOX 识别框架、BroadNet 主干网络和 PAFPN 特征融合网络，表 3.4 给出了分别用 Cascade-CAM、Ultra-SAM 替换 CBAM 方法中的相应模块后的性能表现，从而对比了单个模块的提升效果和整体改进效果。

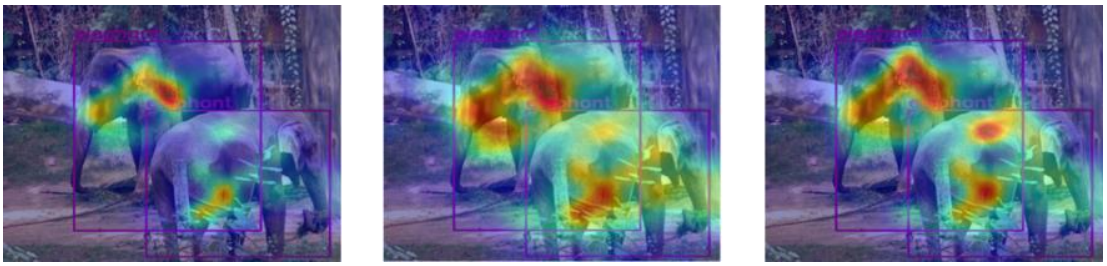
表 3.4 Cascade-CAM、Ultra-SAM 各自对 Ultra-CBAM 的影响

方法	AP50 (%)	mAP (%)
YOLOX_BroadNet_PAFPN+CBAM	75.54	48.11
YOLOX_BroadNet_PAFPN+CBAM_Cascade-CAM	76.02	48.59
YOLOX_BroadNet_PAFPN+CBAM_Ultra-SAM	75.57	48.48
YOLOX_BroadNet_PAFPN+Ultra-CBAM	**76.26**	**48.62**

如表 3.4 所示，Cascade-CAM、Ultra-SAM 都带来了一定程度的性能提升。两者对性能的提升是垂直关系，联合使用可以共同发挥作用带来更好的性能效果。

本节进一步通过热力图可视化的形式定性展现 Cascade-CAM 和 Ultra-SAM 的作用。在同一位置，分别将无注意力引导的特征图（图 3.10（a））、仅通道注意力引导的特征图（图 3.10（b））、通道注意力和空间注意力引导的特征图（图 3.10（c））进行热力图可视化，如图 3.10 所示。颜色越红代表激活程度越大，具有更大的响应值。

原始图像为两头大象，无注意力引导的特征图（a）关注到了大象的耳朵和腿的部分，部分关注到了大象的鼻子，但大象的身体部位存在很大一部分是关注程度较低的。在经过通道注意力 Cascade-CAM 的引导后，更多大象的身体部分被激活，这表明通道注意力在冗余的特征图通道中为含有目标的高区分度通道赋予了更大的权重，使得整体目标区域被更大程度地激活。但由于通道注意力会对特征图的不同空间区域进行同等变换，因此可能也激活了部分非目标区域。



(a) 原始特征图　(b) Cascade-CAM 关注后特征图　(c) Ultra-SAM 关注后特征图

图 3.10 对 PAFPN 中间特征图进行通道注意力关注和空间注意力关注的热力图分析

空间注意力在一定程度上解决了这个问题。如图 3.10（c），在经过 Ultra-SAM 的关注后，整体的红色部分面积减小，但更加聚焦于大象的身体部分，这说明空间注意力为目标区域分配了更大比重，而削弱了非目标区域的比重。

（3）Cascade-CAM 的参数选择

Cascade-CAM 中，一维卷积子网络的卷积核大小由网络自行决定，当输入特征图通道数增多时，卷积核大小也会增大，以保证跨通道信息交互的窗口与总通道数成一定比例关系。由式 3.1，对窗口大小产生影响的主要有两个超参数 γ 和 b。本节希望通过实验找到一种比较好的参数配置。

表 3.5 不同 γ 和 b 下 Ultra-CBAM 的性能表现

γ	b	AP50 (%)	mAP (%)
1	1	75.78	48.02
1	2	76.02	48.59
1	4	75.90	48.37
2	1	75.56	47.95
2	2	76.26	**48.62**
2	4	75.83	48.35
4	1/2/4	**76.52**	48.41

文献[48]中 γ 和 b 的最优取值为 2 和 1，本文据此在 2，1 附近进行取值考虑。为了便于取值，本文限制 $b \le \gamma$，使用 γ 来控制窗口大小。本文实验中特征图层数最大不超过 2048 层，即 C=2048，当 γ=4 时，一维卷积子网络窗口大小为 3，此时再增大 γ 取值便无意义。因此，将参数区间限制在(0, 4]区间内可以获得较为合理的结果。将该区间离散化后，本文枚举了 γ 和 $b \in \{1, 2, 4\}$ 的几种不同情况。

保持 b 不变，γ 越大，一维卷积子网络的卷积核越小。本节给出当 kernel size=7、reduction rate=8 时，Ultra-CBAM 在不同 γ 和 b 取值下的性能表现。如表 3.5 所示，γ=2、b=2 的取值取得了比较均衡的效果。不管是缩小 γ 为 1 还是增大 b 为 4，AP50 和 mAP 大小均有不同程度的下降。这表明盲目增大卷积子网络的窗口大小可能会损害整体的性能。整体而言，γ=b=2 时，性能比较均衡。

（4）Ultra-SAM 中信息聚合方法的选择

在 Ultra-SAM 中的信息聚合阶段和生成空间注意力图时，都涉及到了对卷积层输出、最大池化层输出、平均池化层输出的聚合过程。聚合方式又存在两种选择：通道维度拼接和求和。本节通过实验确定最优的聚合方式搭配。

表 3.6 不同信息聚合方式对 Ultra-SAM 的影响

通道信息聚合方式	空间注意力图生成方式	AP50 (%)	mAP (%)
拼接	拼接	75.09	48.19
拼接	求和	**76.26**	**48.62**
求和	拼接	75.29	48.25
求和	求和	76.22	48.52

由表 3.6 可以看出，用拼接的方式聚合通道信息，用求和方式融合不同空间注意力图可以让 Ultra-SAM 表现出最好的性能。本文在接下来的说明和比较中，均采用“拼接+求和”的方式作为 Ultra-SAM 的设置。

（5）不同 kernel size、reduction rate 下的性能比较

本节讨论空间注意力中卷积核大小 K 和 reduction rate 对性能的影响，希望找到比较合适的参数选择。本文比较了 K∈{3, 7}，reduction rate∈{4, 8, 16}时的不同组合下的性能结果。

表 3.7 Ultra-CBAM 不同 reduction rate 和 kernel size 设置下的性能对比

Kernel size	reduction rate	AP50(%)	mAP (%)	Params (M)	FLOPs (G)
3	4	75.72	48.75	26.33	46.50
3	8	76.19	48.62	26.23	46.38
3	16	**76.51**	**48.90**	**26.19**	**46.33**
7	4	75.34	48.40	26.34	46.52
7	8	76.26	48.62	26.23	46.39
7	16	75.27	47.78	26.19	46.33

如表 3.7 所示，当 kernel size 为 3 时，增大 reduction rate 一定程度上提高了识别效果。reduction rate 为 16 时，AP50 和 mAP 表现最好，mAP 达到了 48.90%。增大卷积核大小并未带来明显的性能提升，当 kernel size 为 7，reduction rate 为 8 时，也有不错的

性能效果，但参数量相比于前者更高，并且卷积核尺度增大也带来了浮点数计算量的增加。

3.4.3 对比实验分析

（1）Ultra-CBAM 与其他注意力机制的比较

本节将本章设计的 Ultra-CBAM 与其他注意力机制进行对比，目的是验证 Ultra-CBAM 在野生动物场景下的适用性。

表 3.8 展示了在 PAFPN 上使用不同的注意力模块的性能比较，包括 AP50 和 mAP 指标，以及参数量和浮点运算量（FLOPs）。这里给出了 SENet、ECANet、CBAM 等注意力机制在 PAFPN 上的表现，每个方法都经过调参，对 kernel size∈{3, 7}，reduction rate∈{8, 16}，以及不同的插入位置上选取了最好性能结果。

表 3.8 在 PAFPN 上，Ultra-CBAM 与其他注意力方法的最好性能对比

注意力方法	AP50 (%)	mAP (%)	Params (M)	FLOPs (G)
-	75.65	48.27	**26.15**	**46.28**
SE-16[47]	75.70	48.51	26.18	46.28
Raw ECANet-11[48]	72.25	46.59	26.15	46.28
Stack ECA-7	75.24	48.31	26.15	46.28
CBAM-3-8[49]	76.21	48.21	26.18	46.28
CBAM-7-8	75.73	48.33	26.23	46.28
Ultra-CBAM-3-16	**76.51**	**48.90**	26.19	46.33

(注：方法名格式“*-7-8”,7 为卷积核大小，8 为 reduction rate)

在无注意力监督的实验中，Wildlife 野生动物数据集的测试集 AP50 为 75.65%，mAP 为 48.27%。Ultra-CBAM-3-16 是空间注意力卷积核尺度为 3，reduction rate 为 16 的 Ultra-CBAM 模块， AP50 和 mAP 最高，带来了 0.86%的 AP 提升以及 0.63%的 mAP 提升，效果最好，实验证明本章设计的 Ultra-CBAM 注意力机制在野生动物场景下优于其他同类注意力机制。

（2）与基准模型的性能比较

本节将本章基于注意力机制的野生动物识别方法与第二章方法和原始 YOLOX 进行整体对比，验证本章方法的有效性。

表 3.9 与基准模型的性能比较

Method	AP50(%)	mAP(%)	Params(M)	FLOPs(G)
YOLOX	68.58	41.58	29.57	54.14
第二章方法	71.80	42.32	**17.70**	**40.09**
本章方法	**76.51**	**48.90**	26.19	46.33

（注:此处给出的参数量为训练时模型的参数量）

如表 3.9，在 Wildlife 测试集上，本章方法 AP50 高于第二章方法 4.71%，mAP 高于第二章方法 6.58%，相应地，参数量增加了 8.49M，浮点数计算量增加了 6.24G，但仍低于原始 YOLOX 模型。与此同时，本章方法 AP 高于 YOLOX 方法 7.93%，mAP 高约 7.32%，证明了本章方法的有效性。

3.4.4 识别结果展示及问题分析

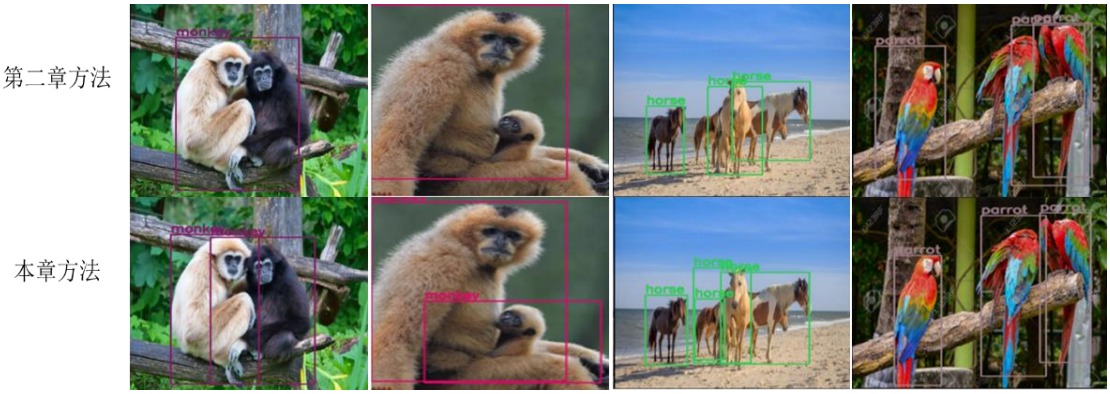


图 3.11 部分样本识别结果展示

图 3.11 展示了本章方法与第二章方法对部分样本的识别结果。最左侧的样本前景目标为两只猿猴，其中褐色猿猴与背景颜色较为接近，并且与白色猿猴部分重叠。第二章方法会漏检此目标，而本章方法可以有效检出。第二列样本为一个大猿猴与一个小猿猴，其中小猿猴的形体大部分被遮挡，且目标较小，姿态不正，属于困难样本。第二章方法只能将大猿猴检出，本章方法可以将小猿猴也识别出来。第三列样本为马群，目标多且遮挡严重，本章方法的识别效果也比第二章更好。最右侧样本为多只鹦鹉，第二章方法对中间鹦鹉定位不准，预测框大面积偏向了右侧鹦鹉，本章方法可以给出更加准确的定位效果。

但正如表 3.9 所示，本章方法的识别效果提升了，但也引入了更多的参数量和浮点数计算量，这对运行所需要的设备的算力资源和运行时空间资源都提出了挑战。在第四章，本文将对本章方法进行模型轻量化。

3.5 本章小结

由于野生动物场景存在背景复杂、目标遮挡等特点，第二章方法对部分样本无法精确定位和识别，存在漏检、定位不准等问题，有进一步改进的空间。本章提出了一种基于注意力机制的野生动物识别方法，基于本章设计的通道-空间联合注意力模块 Ultra-CBAM 构建了特征融合网络 ATT-PAFPN，可以提取更具有区分度的特征图。Ultra-CBAM 使用分层的一维卷积子网络构建了 Cascade-CAM 通道注意力模块来建模通道间的局部相关性，关注不同的野生动物区域。使用分组卷积、平均池化和最大池化来生成空间注意力图并进行融合。本章基于注意力机制的野生动物识别方法可以识别出一些第二章方法无法有效识别的困难样本，整体性能更好。

第四章 基于知识蒸馏的轻量化野生动物识别方法研究

4.1 概述

在第三章，本文阐述了一种基于注意力机制的野生动物识别方法，主要关注如何增强识别模型的性能。但提升性能后的野生动物识别模型存在参数量大、计算量大、推理耗时长等问题，对边缘设备不友好。因此，本章对改进后的模型进行轻量化，目的是减少模型的参数量、计算量，提高运行速度。但仅仅构建并训练一个参数量少的轻量模型必然会带来性能上的损失，可能无法满足对识别准确率的要求。因此，本章关注如何在轻量化的同时尽可能减小模型的性能损失，让轻量化后的模型依然具有不错的识别性能。对这个问题，知识蒸馏是一个很好的解决方案。

知识蒸馏把轻量化模型作为学生，预训练一个高性能大模型作为教师，在训练轻量化模型时同时让教师和数据集真值进行监督，可以提高学生模型的训练效果。对目标识别网络进行知识蒸馏的过程中，存在一些固有问题会影响蒸馏效果，如前景和背景像素之间的不平衡、缺乏对有用区域的关注等。杜治兴等人[65]根据模型分类头的响应生成特征丰富度评分，构造软掩码，为有用区域赋予更大的关注权重。文献[66]提出构造二值化掩码为前景像素赋予更大权重。注意力图可以有效突出更具区分性的前景区域，文献[68]提出，可以分别使用通道注意力池化和空间注意力池化生成教师模型和学生模型中间特征的通道注意力掩码和空间注意力掩码，同时学习教师模型和学生模型掩码和注意力特征图之间的相似性，取得了不错的蒸馏效果。

为了解决第三章改进后的野生动物识别方法参数量大、计算复杂度高、推理耗时长等问题，本章对该方法进行了轻量化，在保留模型结构特色的基础上，构建了一个参数量小、计算量小的轻量化识别模型。由于轻量化后模型精度大幅度降低，本章设计了一种注意力引导的混合蒸馏方法引导轻量化模型的训练过程，提高训练效果。基于第三章设计的注意力模块 Cascade CAM 和 Ultra-SAM。本章设计的混合蒸馏方法根据教师模型和轻量化学生模型的中间特征生成通道注意力掩码和空间注意力掩码，对中间特征的蒸馏过程进行引导，为前景区域分配更大的损失函数比重；同时，基于特征丰富度评分机制对预测输出的蒸馏损失函数进行动态比重分配，通过网络的目标置信度预测对不同像素位置进行评分，可以为重要像素区域分配更高分数优先学习，基于该分数进行预测头的加权蒸馏。在 Wildlife 数据集上，通过本章蒸馏方法训练后的轻量模型识别性能与原始 YOLOX 识别方法相当，但大大降低了参数量和浮点数计算量。本章提出的轻量化野生动物识别模型参数量仅 3.8M，可以实时推理。

4.2 相关工作

4.2.1 野生动物识别轻量化模型设计

本节主要工作为将第三章提出的野生动物识别模型进行轻量化压缩，降低其参数量，轻量化后的识别模型整体结构如图 4.1 所示。本章对第三章识别方法的主干网络、特征融合网络、预测头等都进行了轻量化改造。虽然参数量大大降低了，但每个部分的基本结构和特色都得到了保留。接下来分别介绍模型各部分的轻量化思路。

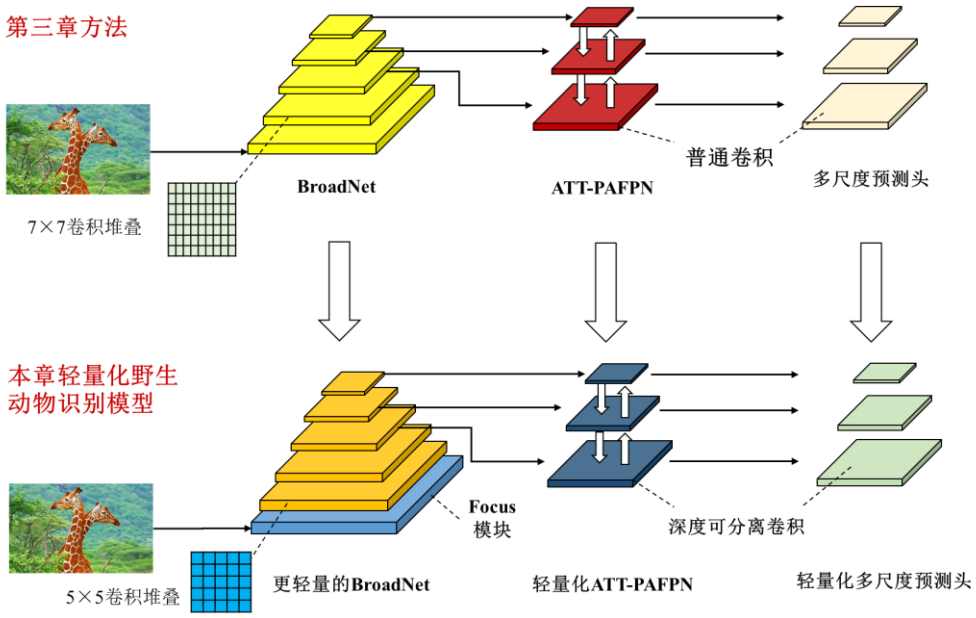


图 4.1 轻量化野生动物识别模型整体结构

对于主干网络的轻量化，本文将 BroadNet 的 kernel size 从 7 降到 5，并且对网络的基本通道数进行减半，同时使用 Focus 模块[27]代替 Stem 模块作为输入层，Focus 模块的结构如图 4.2 所示。Focus 模块将输入特征图在垂直和水平两个方向上每间隔 2 个像素位置进行下采样，得到 4 个尺度为原来一半的特征图，将他们在通道维度上进行拼接，再使用 1×1 卷积进行通道变换。这种方式通过切片实现了下采样，无损地保留原始特征的信息，降低了参数量。

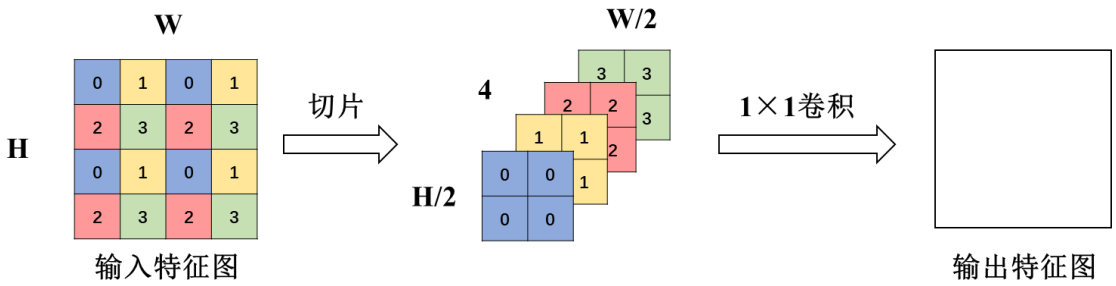


图 4.2 Focus 模块

对于特征融合网络 ATT-PAPFN 和网络的预测头，本章保留其拓扑结构，将每一个

普通卷积使用深度可分离卷积替换。对注意力机制 Ultra-CBAM，由于其本身占用参数量非常小，且对性能提升具有重要作用，因此予以保留。

本节通过实验对比了轻量化前后模型在 Wildlife 测试集上的差距。如表 4.1 所示，第三章方法提出的野生动物识别模型在测试集上具有较高的识别性能，AP50 达到 76.51%，mAP 为 48.90%，但是参数量、浮点数计算量都很大，推理时间单张 416×416 图像平均需要 12.48ms。而对该模型进行轻量化后，推理速度为原来的 4 倍，完全满足实时性的要求，浮点计算量也大幅度下降。但相应地，轻量化模型的 AP50 仅 67.03%，低于轻量化前约 9.5%，mAP 低于轻量化前 11.47%，识别效果大幅度下降。因此，有必要对轻量化模型进行知识蒸馏，从而降低性能损失。

表 4.1 轻量化前后模型对比

方法	AP50 (%)	mAP (%)	Params (M）	FLOPs (G)	推理时间(ms)
第三章方法	**76.51**	**48.90**	26.19	46.33	12.48
本章轻量化模型	67.03	37.43	**3.80**	**6.75**	**3.23**

4.2.2 目标识别领域中的知识蒸馏

对野生动物的监测和识别，通常使用相机陷阱或野外的摄像头等设备。这些边缘设备的算力较低，要求野生动物识别算法既有实时的推理效率，又具有一定的识别性能。对于前者，可以使用更加轻量化的识别网络来达到目的。而对于识别性能的问题，如果直接训练一个轻量化网络，其性能可能无法让人满意，需要使用知识蒸馏这类特殊的训练技巧。

在目标识别领域的知识蒸馏研究中，一些方法采用了同时进行中间特征和预测头输出蒸馏的方式，取得了不错的蒸馏效果，本文称之为混合蒸馏。Chen 等[59]提出了一个端到端训练的多类目标识别混合蒸馏框架，使用加权交叉熵损失函数来进行分类头的蒸馏，使用有界的回归损失函数来进行定位头的蒸馏，并使用过渡适应层来进行中间特征的隐式学习，使学生能够更好地学习教师中间层的输出分布。杜等[65]关注到现有的目标识别蒸馏方法存在的局限性。认为这些方法忽略了边界框之外的有益特征。为了解决上述问题，文献[65]提出使用一种新的特征丰富度评分(Feature Richness Score，FRS)方法来选择重要特征。通过教师模型的分类头输出来生成空间分数掩码，同时进行中间特征的蒸馏和分类定位头的蒸馏，并基于该掩码为不同空间位置的蒸馏损失函数赋予不同权重，提高了蒸馏效果。

一些蒸馏方法使用了注意力机制来提高学生网络的性能。注意力图可以很好地反映不同位置的重要程度[67]，这类方法的核心是在进行中间特征蒸馏时定义基于该特征的注意力图，基于该注意力图的引导在教师和学生之间进行高效的知识迁移。文献[62]通过让学生模型模仿教师模型的注意力图来提升学生模型的性能，把注意力图作为教师模型的“知识”，在学生和教师之间进行注意力图的迁移，整个过程不会引入额外的推

理时参数。文献[68]提出，可以分别使用通道注意力池化和空间注意力池化生成教师模型和学生模型中间特征的通道注意力掩码和空间注意力掩码。一方面学习教师模型和学生模型掩码之间的相似性，另一方面学习基于掩码的注意力特征图之间的相似性。

本文引入第三章提出的 Cascade-CAM 和 Ultra-SAM 注意力模块，对中间特征的蒸馏过程进行引导，同时结合特征丰富度思想生成前景掩码，对预测头的蒸馏过程进行引导，从而提出一种有效的混合蒸馏方法。

4.3 基于知识蒸馏的轻量化野生动物识别方法

直接训练轻量化识别模型不能得到理想的识别效果，存在较大的性能损失，知识蒸馏可以通过教师模型的辅助对轻量化模型的训练过程进行引导，减小性能损失。本节提出一种让学生模型同时学习教师模型的中间特征和预测头输出的混合蒸馏方法。接下来先对整体的蒸馏方法设计进行总述，再分别对基于注意力引导的中间特征蒸馏模块和基于预测头输出的蒸馏模块进行介绍。

4.3.1 整体设计

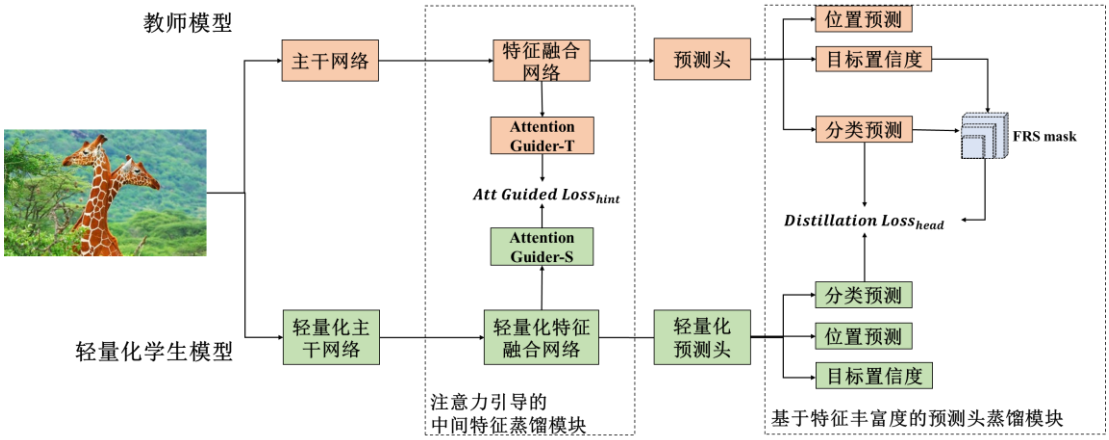


图 4.3 整体蒸馏方法设计

本章整体蒸馏方法设计如图 4.3 所示。把轻量化模型作为学生模型，轻量化之前的模型作为教师模型，总体的蒸馏过程包括两大部分：中间特征蒸馏模块和对最终预测输出的蒸馏模块。在蒸馏训练时，在原有损失函数优化目标的基础上增加两个损失函数：$Loss_{hint}$ 和 $Loss_{head}$，分别用于教师模型对学生模型的特征融合网络输出的三个层级的中间特征图和最终预测头输出的蒸馏，结合训练数据集的标注联合训练。蒸馏过程的总损失函数可以表示为：

$$Loss_{Total} = \alpha Loss_{hint} + \beta Loss_{head} + Loss_{GT} \quad (4.1)$$

α 代表中间特征的蒸馏损失函数的权重，β 代表预测头输出的蒸馏损失函数的权重。$Loss_{GT}$ 代表学生模型对训练数据集标注真值学习产生的损失函数，包括分类头损失函数和定位头损失函数。

为了让学生模型高效地学习教师模型的中间特征，本文为教师模型和学生模型各

设计一个注意力引导模块（Attention Guider），用于突出特征图中更具有区分度的像素区域以及特征通道，对中间特征的蒸馏过程进行引导。与此同时，让学生模型同时学习教师预测头的输出。本文基于文献[65]提出的特征丰富度思想，在其基础上进一步利用目标置信度分数来获得更精确的预测头前景区域掩码。目标置信度分数衡量了一个像素位置是否存在目标。使用交叉熵损失函数度量来计算教师模型和轻量化学生预测头的差异，并根据前景区域掩码重新分配不同像素位置损失函数的比重。

特征丰富度评分可以反映教师模型的真实预测输出，聚合了各个类别的概率分布信息。而注意力则作为一种自适应的前景掩码生成机制，本章方法将两种机制结合使用，从中间特征和预测头输出两个层次提取更精确的前景区域，提高蒸馏的效果。接下来先对基于注意力引导的中间特征蒸馏方法进行阐述。

4.3.2 基于注意力引导的中间特征蒸馏模块设计

注意力引导的中间特征蒸馏方法可以有效突出重要空间位置和重要通道，在蒸馏训练的优化过程中为他们给予更大的权重。如图 4.4 所示，本节方法以教师模型和学生模型的融合网络输出的三个层级 C3、C4、C5 的特征图为输入，计算中间特征蒸馏损失函数 $Loss_{hint}$。

$Loss_{hint}$ 的计算基于教师的注意力引导模块 Attention Guider-T 和学生的注意力引导模块 Attention Guider-S，每个注意力引导模块都包括过渡层（Adaptation Layer）、Ultra-SAM 空间注意力模块和 Cascade-CAM 通道注意力模块。Ultra-SAM 和 Cascade-CAM 分别用于生成空间注意力图和通道注意力向量掩码，他们的具体结构可见第三章相关内容。在生成的教师注意力图和注意力向量掩码的基础上，学生模型进行以下两方面的学习：

1. 让学生模型学习教师模型生成的注意力图和注意力向量掩码。
2. 让学生模型基于掩码学习教师模型的中间特征。计算出各个像素位置的损失函数后，根据注意力图和注意力向量掩码对各位置进行加权，以突出重要像素。

由于学生模型中间特征图的输出通道数与教师模型不一致，无法直接进行损失函数计算，这个问题需要使用过渡层来解决。过渡层使用 1×1 卷积实现，主要作用是对齐教师模型特征图和学生模型特征图的尺度。

总体而言，$Loss_{hint}$ 可表示为：

$$Loss_{\text{hint}} = Loss_{SAM} + Loss_{CA} + Loss_{Feat} \tag{4.2}$$

其中 $Loss_{SAM}$ 和 $Loss_{CA}$ 表示对注意力掩码的学习，$Loss_{Feat}$ 表示对基于掩码对中间特征的加权学习，接下来对他们进行具体介绍。

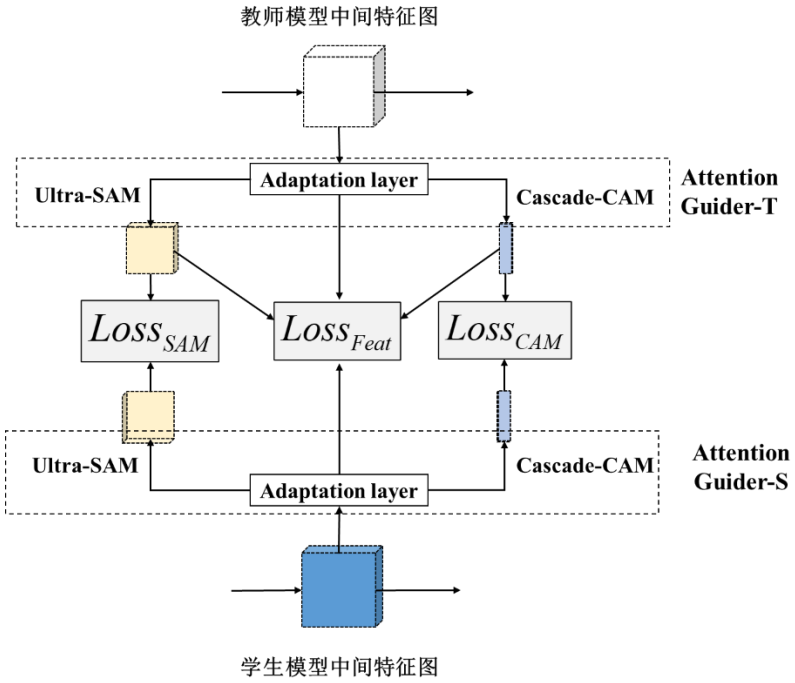


图 4.4 基于注意力引导的中间特征蒸馏方法

（1）对注意力掩码的学习

$Loss_{SAM}$ 是教师模型和学生模型的空间注意力图之间的差异，可表示为：

$$Loss_{SAM}=\sum_{i=1}^{3}CE(SAM_t^i\left(W_{adp_t}^i\left(F_t^i\right)\right),SAM_s^i\left(W_{adp_s}^i\left(F_s^i\right)\right)) \tag{4.3}$$

其中 CE 表示交叉熵损失函数，SAM_t^i、SAM_s^i 分别表示用于生成教师模型和学生模型第 i 个特征层级的空间注意力掩码的 Ultra-SAM 模块。$W_{adp_t}^i$、$W_{adp_s}^i$ 为层级 i 教师模型和学生模型中间特征的过渡层。

$Loss_{CA}$ 是教师模型和学生模型的通道注意力向量之间的差异，可表示为：

$$Loss_{CA}=\sum_{i=1}^{3}CE(CA_t^i\left(W_{adp_t}^i\left(F_t^i\right)\right),CA_s^i\left(W_{adp_s}^i\left(F_s^i\right)\right)) \tag{4.4}$$

其中 CA_t^i、CA_s^i 分别表示用于生成教师模型和学生模型第 i 个特征层级的通道注意力掩码的 Cascade-CAM 模块。

（2）对中间特征的加权学习

$Loss_{Feat}$ 是教师模型和学生模型的中间特征之间的差异，由加权的 L2 损失函数度量，可表示为：

$$Loss_{Feat}=\sum_{i=1}^{3}(W_{adp_t}^i\left(F_t^i\right)-W_{adp_s}^i\left(F_s^i\right))^2\times CA_t^i\left(W_{adp_t}^i\left(F_t^i\right)\right)\times SAM_t^i\left(W_{adp_t}^i\left(F_t^i\right)\right) \tag{4.5}$$

通过教师模型的空间注意力掩码和通道注意力掩码引导，$Loss_{Feat}$ 可以有效对教师模型中间特征图响应的重要空间位置进行关注，突出前景像素特征，优先对教师模型给出的重要区域进行损失函数的优化。接下来阐述基于特征丰富度评分的预测头蒸馏方法。

4.3.3 基于特征丰富度评分的预测头蒸馏模块设计

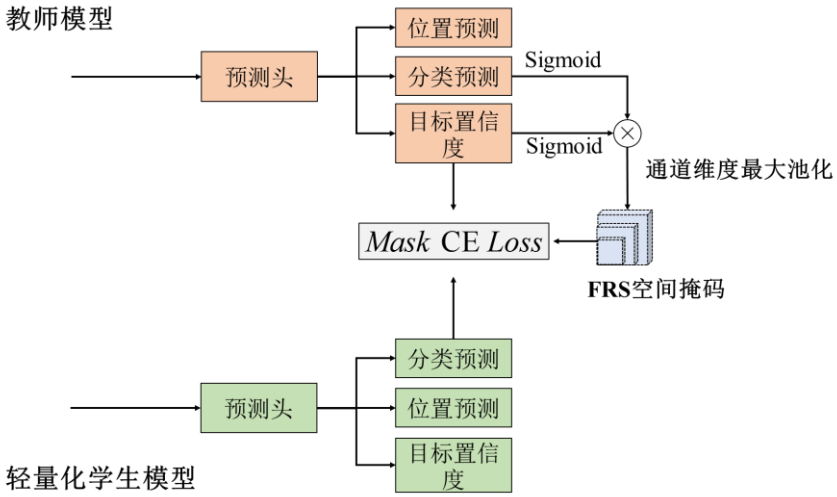


图 4.5 基于特征丰富度评分的分类头蒸馏方法

本节对基于特征丰富度评分的预测头蒸馏模块进行阐述。特征丰富度评分的高低可以反映不同空间位置对分类识别的重要程度。如图 4.5 所示，学生模型和教师模型的预测头包括位置回归预测、类别预测以及目标置信度预测。本节方法只对学生模型的分类头进行蒸馏。但在蒸馏时，本文同时利用分类头输出的概率分布和目标置信度预测来辅助生成特征丰富度。

某空间位置的特征丰富度评分包括两部分，一是将该位置的分类头输出使用 Sigmoid 函数激活，得到 0-1 内的概率分布，通过在不同类间求取最大值对该分布信息进行聚合；二是目标置信度预测输出的分数。目标置信度预测输出的分数也包含了一定的语义信息，反映了模型对该位置含有目标的信心，可以用来辅助特征丰富度的生成过程。将两者相乘即可。给定特征 F ，本文的特征丰富度可描述为：

$$FRS = \max_{1 \le c' \le C} \left(P(c' \mid F) \right) \times P(obj \mid F) \tag{4.6}$$

其中 C 为类别总数，$P(c' \mid F)$ 表示某类别 c' 的置信概率，$P(obj \mid F)$ 表示目标置信度分数。进而，基于特征丰富度评分的预测头蒸馏损失函数 $Loss_{head}$ 可以由交叉熵计算为：

$$Loss_{head} = \sum_{i=1}^{3} CE(O_t^i, O_s^i) * FRS_t^i \tag{4.7}$$

式中 CE 表示交叉熵损失函数，O_s^i、O_t^i 分别表示学生模型和教师模型第 i 个层级的分类头输出。FRS_t^i 表示基于教师模型的预测输出计算得到的层级 i 的特征丰富度掩码。对每一个空间位置都进行特征丰富度评分，可以得到教师模型的特征丰富度掩码（Feature Richness Score Mask, FRS Mask）。进一步根据该掩码对分类头的蒸馏过程进行增强，可以突出有目标的区域，对前景像素优先进行蒸馏。

4.4 实验与分析

本节通过实验验证本章提出的轻量化野生动物识别模型以及混合蒸馏方法在野生动物场景下的有效性。主要包括以下几个方面：

1. 通过消融实验对比和特征丰富度蒸馏模块的作用。
2. 对中间特征蒸馏模块中不同注意力引导机制的作用进行消融分析。
3. 对比本章蒸馏方法与其他蒸馏方法在野生动物场景下的蒸馏效果。
4. 对比蒸馏后的轻量化模型与基准模型 YOLOX 的目标识别效果。

4.4.1 实验基础设置

本节实验中，学生模型使用 4.2 节提出的轻量化野生动物识别模型，教师模型为轻量化前的识别模型。评价指标采用 mAP 和 AP50。教师模型在 Wildlife 野生动物数据集上进行了预训练，以在测试集性能上最好的训练轮次的对应模型参数作为教师模型的权重。在学生模型的训练过程中，数据增强策略以及 Batch size 等超参数与前文保持一致。

4.4.2 消融实验分析

本节通过消融实验分析对蒸馏效果可能产生影响的因素。包括基于中间特征蒸馏和基于特征丰富度蒸馏两个蒸馏模块的作用、基于中间特征的蒸馏模块中空间注意力和通道注意力的作用。

（1）各蒸馏模块的作用

本节对注意力引导的中间特征蒸馏模块和特征丰富度蒸馏模块进行消融分析。目的是对各个蒸馏模块的效果进行对比。如表 4.2 所示，本节对比了不使用蒸馏直接训练的轻量化学生模型、只使用注意力引导蒸馏模块、只使用特征丰富度蒸馏模块以及同时使用注意力引导模块和特征丰富度蒸馏模块四种设置下的识别效果。

表 4.2 各蒸馏模块对蒸馏效果的影响

注意力引导蒸馏模块	特征丰富度蒸馏模块	AP50 (%)	mAP (%)
×	×	67.03	37.43
√	×	68.95	38.98
×	√	69.66	39.52
√	√	**71.28**	**40.52**

由表中可以看出，注意力引导的中间特征蒸馏可以为轻量化识别模型带来 1.92%的 AP 提升和 1.55%的 mAP 提升，而特征丰富度蒸馏模块可以为轻量化识别模型带来 2.63%的 AP 提升和 2.08%的 mAP 提升。整体而言，两个蒸馏模块对最终蒸馏效果都具有重要作用。同时使用带来了较大的性能提升。

（2）不同注意力引导模块的作用

本节对注意力引导蒸馏模块中的空间注意力机制和通道注意力机制进行消融分析，在实验时，只使用空间注意力引导的损失函数和通道注意力引导的损失函数。

表 4.3 各注意力损失函数对蒸馏效果的影响

Channel Attention Loss	Spatial Attention Loss	AP50 (%)	mAP (%)
×	×	67.03	37.43
√	×	67.88	38.09
×	√	68.23	38.56
√	√	**68.95**	**38.98**

如表 4.3 所示，在中间特征蒸馏时，加入通道注意力引导损失函数带来了约 0.85%的 AP 提升以及约 0.66%的 mAP 提升。空间注意力损失函数的增益效果更明显，分别带来了 1.2%的 AP 提升以及 1.13%的 mAP 提升。两者同时使用取得了更好的优化效果。

4.4.3 对比实验分析

本节进行两个层次的实验对比，一是对比本章提出的基于注意力引导的混合蒸馏方法与其他蒸馏方法的蒸馏效果进行对比，目的是验证本章方法在野生动物场景下的效果；二是把经过轻量化和本章方法蒸馏后的模型与全文基准模型原始 YOLOX 在 Wildlife 数据集上的性能、参数量、计算代价等进行对比。

（1）与其他蒸馏方法的对比

本小节给出在相同的教师模型和学生模型设置下，几种不同的蒸馏方法为轻量化学生模型带来的性能增益。

表 4.4 与其他蒸馏方法的对比

Distillation Method	AP50 (%)	mAP (%)
不使用蒸馏	67.03	37.43
Hinton[51]	67.75	37.99
FitNets[59]	67.43	37.91
FRS[65]	70.62	40.42
本章蒸馏方法	**71.28**	**40.52**

(注:“-”表示不使用蒸馏训练时的学生模型)

如表 4.4 所示，如果不使用知识蒸馏，而是对轻量级学生模型直接进行训练，在 Wildlife 测试集上，AP50 只有 67.03%，而 mAP 只有 37.43%。使用 FitNet 这类完全基于中间特征的蒸馏方法为学生模型带来了约 0.4%的 AP 和 mAP 提升。Hinton[51]这类基于最终预测输出的蒸馏方法也提升了约 0.72%的 AP50 以及 0.56%的 mAP。FRS[65]基于

特征丰富度评分掩码对中间特征和最终预测输出进行混合蒸馏，带来了 3.59%的 AP 提升。本章方法为轻量化模型带来了 4.25%AP 提升以及 3.08%mAP 提升，具有更好的蒸馏效果。

（2）与其他识别模型的对比

本章的轻量化识别模型保留了第二章基于主干网络的改进、第三章基于注意力机制的改进，并使用本章设计的知识蒸馏方法提升了训练效果。因而本章轻量化模型综合利用了本文各章的改进方法。有必要将其与改进前的原始YOLOX模型进行对比，以评估本文的整体工作效果。

表 4.5 与全文基准识别模型 YOLOX 的对比

方法	AP50(%)	mAP(%)	Params(M）	FLOPs(G)	推理时间 (ms)
YOLOX	68.58	41.58	29.57	54.14	9.22
YOLOv4[25]	67.82	40.22	41.19	29.56	11.02
YOLOv5[27]	68.27	40.64	21.09	21.42	8.47
FCOS[22]	71.21	**42.27**	32.13	68.36	20.19
蒸馏后的 轻量化识别模型	**71.28**	40.52	**3.80**	**6.75**	**3.23**

如表 4.5 所示，以 CSPDarkNet 作为主干网络，FPN 作为特征融合网络，YOLOX 识别模型在 Wildlife 数据集上 mAP 为 41.58%，AP50 为 68.58%。其参数量约 29.57M，浮点数运算量达到 29.28G，在本文实验使用的英伟达 RTX 3060GPU 上，推理一个 416×416 分辨率的单张图片平均时长约为 9.22ms。

相比之下，蒸馏后的轻量化野生动物识别模型 AP50 高出 YOLOX 模型约 2.7%，mAP 低约 1.06%，性能基本相当。但参数量只有 3.8M，减少了约 87.1%，浮点数计算量只有 6.75G，减小了 87.6%，推理耗时平均为 3.23ms，是基准模型 YOLOX 的 35%。

本章的轻量化模型的参数量、计算量、推理时间也小于如 YOLOv4、YOLOv5 等其他目标识别方法，在性能上虽然 mAP 略有劣势，但 AP50 上具有优势。FCOS 具有最高的 mAP，但该方法计算复杂度高，参数量也大幅度高于本章轻量化模型。整体而言，蒸馏后的轻量化野生动物识别模型兼顾了识别性能和推理效率，达到了本文的研究目标。

4.4.4 识别结果展示

本节给出轻量化模型的部分可视化识别结果，如图 4.6 所示。轻量化模型对单目标、多目标、遮挡样本等都有不错的识别效果。对于背景复杂、前背景易混淆的样本，如图 4.6 中的猴子样本和雪兔样本，也有一定的识别准确率。

图 4.6 轻量化识别模型部分可视化结果

4.5 本章小结

本章主要介绍了一种针对野生动物识别任务的模型轻量化和知识蒸馏方法。通过使用 Focus 模块、深度可分离卷积以及减小通道宽度等方式，可以有效减小模型的参数量和计算量，从而满足边缘设备的资源限制和实时性要求。但是，轻量化过程大幅度降低了识别效果。因此，本章提出了一种混合知识蒸馏方法，包括注意力引导的中间特征蒸馏模块以及基于特征丰富度分数的预测头蒸馏模块。这种蒸馏方法可以在保持轻量化模型推理效率的同时，减少性能损失，提高识别性能。实验结果表明，经过本章方法蒸馏后的轻量化模型，在 Wildlife 数据集上表现良好，与其他目标识别方法相比，识别性能基本持平，但参数量和浮点数计算量大幅度减少，推理快，可以满足实时性的要求。本章的轻量化野生动物模型兼具推理效率和识别性能，达到了本文的研究目标。

第五章 总结及展望

5.1 全文总结

本文经过递进式的改进，最终提出了一种新颖的轻量化野生动物识别方法。基于通用目标识别网络 YOLOX，本文先后对其主干网络、特征融合网络进行了改进，分别在第二章和第三章提出了一种基于结构重参数主干网络的野生动物识别方法和一种基于注意力机制的野生动物识别方法，大大提高了 YOLOX 在野生动物识别任务上的性能。由于提升性能的过程中引入了一些参数量和计算量，可能不利于边缘设备的部署，本文在第四章进行了模型轻量化，并提出一种混合知识蒸馏方法，减少模型轻量化过程中的性能损失。最终提出一个性能与原始YOLOX在野生动物场景下不相上下，但参数量和计算量大幅度降低的轻量化野生动物识别模型，兼顾效率和性能。总体而言，本文的主要工作如下：

（1）提出了一种基于结构重参数主干网络的高效野生动物识别方法

由于 YOLOX 的主干网络特征 CSPDarkNet 在野生动物场景下，特征提取能力不强，有效感受野较小，对某些场景如目标不全或遮挡情况下的野生动物样本无法有效识别。本文在第二章提出一种基于结构重参数主干网络的野生动物识别方法。该方法基于 7×7 的大卷积核设计了特征提取主干网络 BroadNet，并通过深度卷积块来减小大卷积核所带来的计算量。在训练时，使用多分支结构提高网络的表达能力，推理时转化为简洁的直筒式结构，减少了推理时参数，提高了网络的感受野。通过增大感受野，野生动物识别模型可以有效识别一些YOLOX漏检、识别不准的困难样本。基于该主干网络，第二章方法在野生动物数据集 Wildlife 上的整体性能超过了 YOLOX、YOLOv4、YOLOv5 等识别方法。

（2）提出了一种基于注意力机制的高效野生动物识别方法

本文在第三章提出了一种基于注意力机制的野生动物识别方法，旨在解决第二章方法在复杂场景下部分野生动物样本定位不准、漏检等问题。基于特征融合网络 PAFPN 以及本文设计的联合注意力模块 Ultra-CBAM，本文选取了最优的注意力位置来进一步提高在野生动物场景下的识别效果。Ultra-CBAM 综合了 CBAM 的优势，并加以优化，基于一维卷积子网络，提高了通道注意力跨通道的信息交互能力，同时使用参数化的方式优化了空间注意力聚合通道信息的过程。基于 Ultra-CBAM，第三章野生动物识别方法提高了特征图上前景目标区域的响应，继而提高模型对野生动物的识别准确率。通过在 Wildlife 测试集上的实验，第三章基于注意力机制的野生动物识别方法可以识别出一些第二章方法无法有效识别的困难样本，整体性能更好，但相应地引入了额外参数量和计算量。

（3）提出了一种基于知识蒸馏的轻量化野生动物识别方法

为了解决第三章方法可能在边缘设备上存在的计算资源不足和实时性要求问题。本文在第四章提出了一种基于知识蒸馏的轻量化野生动物识别方法。主要包括对第三章方法的轻量化以及通过知识蒸馏减少轻量化过程中的性能损失两部分工作。第四章通过使用 Focus 模块、深度可分离卷积以及减小通道宽度等方式，有效减小了模型的参数量和计算量，从而满足边缘设备的资源限制和实时性要求。但是，这样做会导致识别效果的降低。因此，第四章进一步提出了一种混合知识蒸馏方法来减少性能损失。

混合蒸馏方法包括基于注意力引导的中间特征蒸馏模块以及基于特征丰富度分数的预测头蒸馏模块。这种方法可以在蒸馏过程中为不同的空间和通道位置的蒸馏损失函数动态赋予不同权重，有效突出前景目标，对重要区域优先进行蒸馏。

实验结果表明，蒸馏后的轻量化模型，在 Wildlife 数据集上表现良好，与其他目标识别方法相比，识别性能基本持平，但参数量和浮点数计算量大幅度减少，推理快，可以满足实时性的要求。第四章提出的轻量化野生动物模型兼具推理效率和识别性能，达到了本文的研究目标。

这三部分工作是递进关系，第二章方法关注对主干网络的改进，通过增大感受野提高了识别网络的性能，第三章方法关注对特征融合网络的改进，在第二章方法的基础上通过引入注意力机制等方式，进一步提高网络的性能。第四章对第三章方法进行了轻量化，保留了第二章和第三章中的结构改进，大幅度提高了模型的推理效率，并进一步提出一种混合知识蒸馏方法减少轻量化模型的性能损失。最终，本文提出了一种轻量化野生动物识别模型，其性能与原始 YOLOX 相当，但参数量和计算量大大降低。

5.2 未来展望

本文工作主要围绕算法层面，针对野生动物场景下存在的一些特点和基准模型的问题进行改进。虽然最终提出的野生动物识别模型大大降低了基准模型的参数量，同时有不错的性能，对边缘设备的部署比较友好。但由于资源、硬件等原因，本文并未将算法在设备上进行实际部署。因此，在未来本文希望结合具体硬件平台，做一些涉及到部署推理、算子优化等方面的工作，进一步优化本文方法。

致谢

在本论文即将结束之际，我想借此机会向在我研究生学习道路上给予支持和帮助的人们表示最真诚的感激和衷心的谢意。

感谢学校和相关部门对我学习和生活的关怀和支持，在苏州校区的生活非常愉悦，对我的科研和个人成长起了重要作用。

衷心感谢我的导师路小波教授。路老师知识面广博，治学态度严谨，负责耐心，每次指导总能给出高屋建瓴的意见，让我不断进步和成长。学生朽木，让您费心了。

衷心感谢我的家人、朋友、实验室同门们的支持和鼓励。没有你们的支持，我无法完成我的研究生学业，谢谢你们的陪伴和信任。

参考文献

[1]. 张雪莹,张浩林,韩莹莹,翁强,袁峥嵘,姚远.基于深度学习的野生动物监测与识别研究进展[J].野生动物学报,2022,43(01):251-258.

[2]. 程浙安. 基于深度卷积神经网络的内蒙古地区陆生野生动物自动识别[D]. 北京: 北京林业大学, 2019.

[3]. 王梦来, 李想, 陈奇, 等. 基于 CNN 的监控视频事件检测[J]. 自动化学报, 2016, 42(6): 892-903.

[4]. 何嘉. 基于深度学习的野生动物智能检测与识别[D].深圳大学,2019.

[5]. 蔡前舟,郑伯川,曾祥银,侯金.结合长尾数据解决方法的野生动物目标检测[J].计算机应用,2022,42(04):1284-1291.

[6]. 韩家臣. 基于深度学习的野生动物图像识别方法研究[D].西北师范大学,2021.

[7]. 杨铭伦,张旭,郭颖等.基于 YOLOv5 的红外相机野生动物图像识别[J].激光与光电子学进展,2022,59(12):382-390.

[8]. 徐其森. 基于深度学习的野生动物目标检测与跟踪[D].东北林业大学,2022.

[9]. 王旭,罗铁坚,杨林.基于 Transformer 的野生动物关键点检测[J].传感器世界,2021,27(11):19-25.

[10]. 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.

[11]. Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[12]. Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(06): 1137-1149.

[13]. Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[14]. Uijlings J R R, Van De Sande K E A, Gevers T, et al. Selective search for object recognition[J]. International journal of computer vision, 2013, 104: 154-171.

[15]. Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[16]. Liu L, Ouyang W, Wang X, et al. Deep learning for generic object detection: A survey[J]. International journal of computer vision, 2020, 128: 261-318.

[17]. Cai Z, Vasconcelos N. Cascade r-cnn: Delving into high quality object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6154-6162.

[18]. Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.

[19]. Duan K, Bai S, Xie L, et al. Centernet: Keypoint triplets for object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6569-6578.

[20]. Law H, Deng J. Cornernet: Detecting objects as paired keypoints[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 734-750.

[21]. 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.

[22]. Tian Z, Shen C, Chen H, et al. Fcos: Fully convolutional one-stage object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 9627-9636.

[23]. Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[24]. Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[25]. Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[26]. Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021[J]. arXiv preprint arXiv:2107.08430, 2021.

[27]. Ultralytics. YOLOv5[CP]. 2020. https://github.com/ultralytics/yolov5.

[28]. Lin T Y, Dollár P, Girshick R, et al. [28]mid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[29]. Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014: 740-755.

[30]. Wang C Y, Liao H Y M, Wu Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020: 390-391.

[31]. Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8759-8768.

[32]. Luo W, Li Y, Urtasun R, et al. Understanding the effective receptive field in deep convolutional neural networks[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016: 4905-4913.

[33]. 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.

[34]. 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.

[35]. Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning[J]. arXiv e-prints, 2016: arXiv: 1602.07261.

[36]. 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.

[37]. 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.

[38]. Ding X, Guo Y, Ding G, et al. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1911-1920.

[39]. Ding X, Zhang X, Han J, et al. Diverse branch block: Building a convolution as an inception-like unit[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 10886-10895.

[40]. Ding X, Zhang X, Ma N, et al. Repvgg: Making vgg-style convnets great again[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13733-13742.

[41]. Ding X, Zhang X, Han J, et al. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 11963-11975.

[42]. 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.

[43]. 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.

[44]. Liu S, Chen T, Chen X, et al. More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity[J]. arXiv preprint arXiv:2207.03620, 2022.

[45]. 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.

[46]. 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.

[47]. 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.

[48]. 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.

[49]. Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the

European conference on computer vision (ECCV). 2018: 3-19.

[50]. Park J, Woo S, Lee J Y, et al. Bam: Bottleneck attention module[J]. arXiv preprint arXiv:1807.06514, 2018.

[51]. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.

[52]. Gou J, Yu B, Maybank S J, et al. Knowledge distillation: A survey[J]. International Journal of Computer Vision, 2021, 129: 1789-1819.

[53]. Meng Z, Li J, Zhao Y, et al. Conditional teacher-student learning[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 6445-6449.

[54]. Chen D, Mei J P, Zhang Y, et al. Cross-layer distillation with semantic calibration[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(8): 7028-7036.

[55]. Kim J, Park S U, Kwak N. Paraphrasing Complex Network: Network Compression via Factor Transfer[J]. arXiv preprint arXiv:1802.04977, 2018.

[56]. Jin X, Peng B, Wu Y, et al. Knowledge distillation via route constrained optimization[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 1345-1354.

[57]. Passalis N, Tefas A. Learning deep representations with probabilistic knowledge transfer[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 268-284.

[58]. Müller R, Kornblith S, Hinton G. When Does Label Smoothing Help?[J]. arXiv e-prints, 2019: arXiv: 1906.02629.

[59]. Chen G, Choi W, Yu X, et al. Learning efficient object detection models with knowledge distillation[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 742-751.

[60]. Romero A, Ballas N, Kahou S E, et al. Fitnets: Hints for thin deep nets[J]. arXiv preprint arXiv:1412.6550, 2014.

[61]. Huang Z, Wang N. Like what you like: Knowledge distill via neuron selectivity transfer[J]. arXiv preprint arXiv:1707.01219, 2017.

[62]. 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.

[63]. 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.

[64]. Ma N, Zhang X, Zheng H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 116-131.

[65]. Zhixing D, Zhang R, Chang M, et al. Distilling object detectors with feature richness[J]. Advances in Neural Information Processing Systems, 2021, 34: 5213-5224.

[66]. Wang T, Yuan L, Zhang X, et al. Distilling object detectors with fine-grained feature imitation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4933-4942.

[67]. Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2921-2929.

[68]. Zhang L, Ma K. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors[C]//International Conference on Learning Representations. 2021:1-14

作者简介

吕玉鑫，男，24 岁，山东聊城人，主要研究方向为目标检测。

硕士期间的学术成果：

[1]. 吕玉鑫，路小波. 一种基于 FPN 的注意力机制改进算法.东南大学校庆报告会.2022.

[2]. 吕玉鑫，路小波. An Improved YOLOX Algorithm based on Structural Re-parameterized CBAM for Wild Animals Detection. 4th International Symposium on Smart and Healthy Cities. 2022.