Modeling and Simulation 建模与仿真, 2024, 13(3), 3889-3898
Published Online May 2024 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2024.133354

Hans汉斯

基于YOLOv5的高效轻型化小目标检测模型YOLO-QCK

黄鹏辉，赵　政，李　蓉

上海理工大学计算机与信息工程学院，上海

收稿日期：2024年4月27日；录用日期：2024年5月23日；发布日期：2024年5月31日

摘　要

目标检测是计算机视觉的一个重要方面，在准确性和鲁棒性方面取得了重大进展。尽管取得了这些进步，但实际应用仍然面临着显著的挑战，主要是小物体的检测不准确或漏检。此外，检测模型的大量参数计数和计算需求阻碍了它们在资源有限的设备上的部署。本文提出了一种基于YOLOv5的高级目标检测模型YOLO-QCK。我们首先在颈部网络金字塔结构中引入一个额外的小物体检测层，从而产生一个更大尺度的特征图，以识别小物体的更精细的特征。此外，我们将C3CrossCovn模块集成到骨干网中。该模块采用滑动窗口特征提取，有效地减少了计算量和参数数量，使模型更加紧凑。与基线YOLOv5s模型相比，我们新开发的模型YOLO-QCK在MS COCO验证数据集上显示出相当大的改进，mAP@0.5增加了4.6%，mAP@0.5:0.95增加了4%，同时保持模型尺寸紧凑，参数为9.49 M。结果验证了YOLO-QCK模型在小目标检测中的高效性能，以较少的参数和计算量实现了高精度。

关键词

目标检测，深度学习，YOLOv5s，神经网络

YOLO-QCK: An Efficient and Lightweight Small Object Detection Model Based on YOLOv5

Penghui Huang, Zheng Zhao, Rong Li

School of Computer and Information Engineering, Shanghai University of Technology, Shanghai

Received: Apr. 27th, 2024; accepted: May. 23rd, 2024; published: May. 31st, 2024

文章引用: 黄鹏辉, 赵政, 李蓉. 基于 YOLOv5 的高效轻型化小目标检测模型 YOLO-QCK [J]. 建模与仿真, 2024, 13(3): 3889-3898. DOI: 10.12677/mos.2024.133354

Abstract

Object detection is an important aspect of computer vision, with significant advancements made in accuracy and robustness. Despite these advancements, practical applications still face significant challenges, primarily due to inaccurate detection or missed detection of small objects. Additionally, the large parameter counts and computational requirements of detection models hinder their deployment on resource-limited devices. This paper proposes an advanced object detection model, YOLO-QCK, based on YOLOv5. We first introduce an additional small object detection layer within the neck network pyramid structure to generate a larger-scale feature map for identifying finer features of small objects. Furthermore, we integrate the C3CrossCovn module into the backbone network, which employs sliding window feature extraction to effectively reduce computational complexity and parameter count, resulting in a more compact model. Compared to the baseline YOLOv5s model, our newly developed YOLO-QCK model demonstrates significant improvements on the MS COCO validation dataset, with an increase of 4.6% in mAP@0.5 and 4% in mAP@0.5:0.95, while maintaining a compact model size with 9.49 M parameters. These results validate the efficient performance of the YOLO-QCK model in small object detection, achieving high precision with fewer parameters and computational requirements.

Keywords

Object Detection, Deep Learning, YOLOv5s, Neural Network

1. 引言

近年来，深度学习的快速发展使得计算机视觉的各个方面都取得了重大突破，尤其是在目标检测方面。计算机视觉旨在识别和分类图像中的物体(例如，行人，动物，车辆)，作为目标跟踪和目标分割等任务的基础元素[1] [2]。它的工业应用非常广泛，从探伤到自动驾驶[3] [4] [5]。此外，电子控制系统和飞机设计的发展突出了基于无人机(UAV)的目标检测的重要性，在农业，管理和航空摄影领域非常普遍。无人机要么无线电控制，要么在预先编程的路线上操作。无人机配备高分辨率摄像头，能够捕获全面的数字图像，便于使用轻型模型在飞行中实时检测目标[6] [7] [8]。目前，实现目标检测模型的方法主要有两种：两阶段方法[9] [10]和单阶段方法[11]。

Girshick 等人开发的 R-CNN [12]使用预训练好的 CNN 在各个候选区域进行特征提取，并使用 svm 进行分类。R-CNN 虽然具有创新性，但处理过程冗余复杂，消耗了大量的计算资源和存储空间。Ren 等人提出了 Faster R-CNN [13]，这是目标检测中的一个里程碑模型。更快的 R-CNN 创新地将区域建议生成与特征提取网络相结合，实现了涵盖区域建议、特征提取、分类和边界盒回归的全面端到端训练。Joseph 等人提出的 YOLO [14]通过将输入图像初始划分为几个网格来实现目标检测。每个网格预测对象的潜在边界框，跨越整个图像。然后通过 NMS 对这些预测进行细化，得出最终的检测结果。由 Redmon 等人提出的 YOLOv3 [15]标志着一个重大的进步，用 DarkNet53 取代了它的特征提取网络。YOLOv3 还利用了特征金字塔网络(FPNet) [16]，使用多尺度下采样特征图来增强对小物体的检测，从而提高了推理速度和准确性。Bochkovskiy 等人介绍了 YOLOv4 [17]，该检测模型将其骨干网升级为 CSPDarknet53，这是

DarkNet53 的增强版本，具有 CSP 模块。这种增强使网络更加模块化，简化了骨干网络架构，增强了特征提取能力。YOLOv4 在 YOLOv3 的基础上进一步发展，在颈部增加了路径聚合网络(PAN)架构，借鉴了 PANet [18]的原理，增强了多尺度特征融合。另一方面，Dosovitskiy 等人利用 Vision Transformer (ViT)率先将 Transformer 应用于计算机视觉[19]，彻底改变了目标检测的研究领域。ViT 采用编码器－解码器和自关注机制来捕获全局图像特征，从而实现完全的端到端检测。然而，这种创新的代价是增加了模型参数，增加了培训和部署的挑战。同时，Carion 等人引入了 DETR [20]，首次将 Transformer 架构应用于目标检测。DETR 由基于 cnn 的骨干网、编码器、解码器和前馈网络组成，形成了一个完整的检测系统。尽管它的高参数计数和有限的精度增强，DETR 已成为基于变压器的目标检测的后续进展的基石。基于 cnn 的目标检测方法通常利用深度和广度骨干网络进行特征提取。他们利用多尺度特征融合[21]在不忽略几何纹理细节的情况下捕获广泛的语义信息，从而增强检测特征图的表达能力。然而，这些方法有一个权衡，因为大量的卷积操作和堆叠特征提取网络显着增加了模型复杂性和参数计数。

在本研究中，我们提出了 YOLOv5 的改进版 YOLO-QCK 模型，以解决上述挑战。关键的改进在颈部网络中加强多尺度特征融合以更准确地检测小物体。然而，这些改进可能导致参数数量和计算需求的上升。为此，我们提出并评估了轻量级策略应用于 YOLO-QCK，以确保模型的高效性。

本文的主要贡献如下：

(1) 在 YOLOv5s 模型的颈部网络中加入微小检测层，增强其微小目标检测性能。

(2) 将 C3CrossCovn 模块嵌入到 YOLOv5s 模型的骨干网中，通过减少模型参数和计算量来简化模型。

2. 模型

2.1. YOLO-QCK 整体框架

在本研究中，我们提出了一种基于 YOLOv5 改进的目标检测模型 YOLO-QCK，其重点是小目标检测，降低了模型复杂度，如图 1 所示。YOLOv5 模型分为三个主要部分：骨干网络、颈部网络和头部网络。骨干网络建立在 CSPDarknet53 上，由标准卷积层和附加的特征增强模块组成，任务是提取物体的形状和颜色等几何纹理特征。为了丰富这一基础信息，颈部网络从 FPNet [16]和 PANet [18]中汲取灵感，进一步将骨干网的特征图与更深层次的语义信息相结合。这种组合产生了具有丰富语义和几何信息的特征映射。这些增强的特征图然后被输入到头部网络中，头部网络执行最终的检测和分类。

2.2. 微小目标检测层

在 MS COCO 数据集中，对象根据大小进行分类：小对象小于 32 × 32 像素，中等对象在 32 × 32 到 96 × 96 像素之间，大对象超过 96 × 96 像素。为了增强模型中的小目标检测，我们调整了特征映射和锚点的大小。利用 k-means 聚类算法，重新校准预设锚盒的大小范围，引入微小目标检测层，建立 YOLOv5-sm 模型。具体来说，我们在 YOLOv5s 的颈部网络中进行上采样，生成一个包含 128 个通道的 160 × 160 特征图，然后将其与骨干网络的第三层输出相结合，在通道和大小上进行匹配。该组合特征图与其他检测层输出一起，在头部网络中进行分类和检测处理。我们的颈部网络生成了几个不同尺度的特征图，每个特征图对应不同的锚大小。这样可以在较大的特征图上检测较小的对象，在较小的特征图上检测较大的对象，从而增强图像中的对象表示。特征图大小与锚点大小之间的关系详见表 1。

2.3. 轻量级卷积模块

为了简化 YOLOv5 模型，减少参数数量和计算量，本研究探讨了 C3Ghost 和 C3CrossCovn 两个模块的集成。我们的目标是在前入骨干网内的不同位置时评估这些模块的影响。这种比较侧重于每个模块放

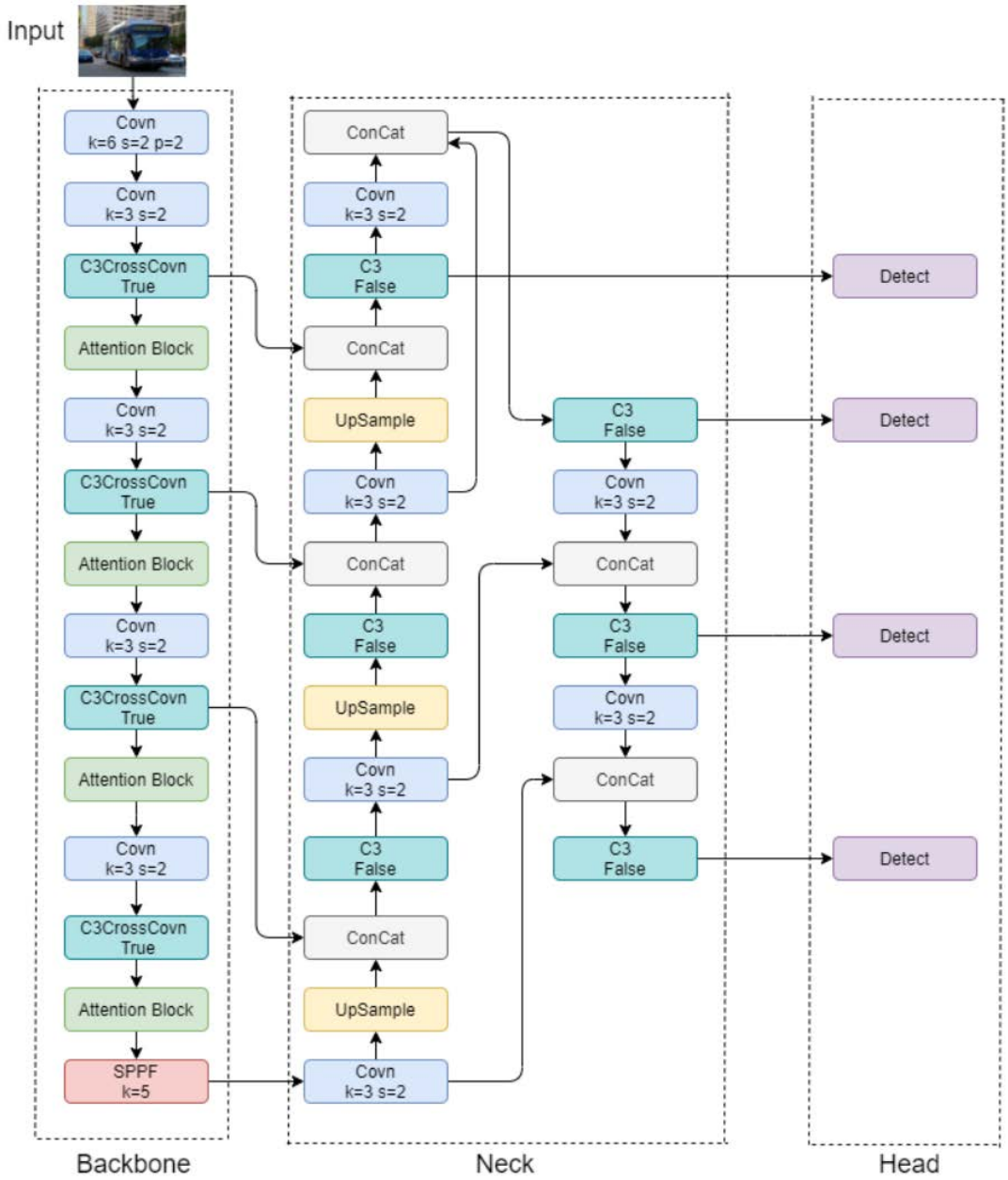

Figure 1. YOLO-QCK network architecture
图 1. YOLO-QCK 网络架构

Table 1. Anchor box sizes in different feature maps
表 1. 不同特征映射中的锚框大小

特征图大小	锚框大小
160 × 160	9 × 12, 20 × 19, 17 × 42
80 × 80	43 × 26, 36 × 56, 76 × 52
40 × 40	49 × 121, 108 × 102, 111 × 121
20 × 20	231 × 138, 230 × 325, 479 × 372

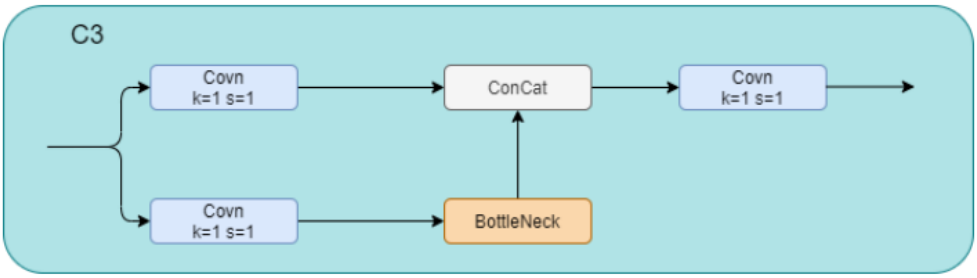

Figure 2. C3 module (k and s indicate the convolutional kernel size and stride, respectively)
图 2. C3 模块(k 和 s 分别表示卷积核的大小和步幅)

置如何影响模型的性能和复杂性，其目标是在不影响其性能的情况下简化网络体系结构。

C3 模块：C3 模块是 YOLOv5 的关键组件，如图 2 所示，由三个标准卷积层组成，每个层的内核大小为 1 × 1，步幅为 1，并且包含多个堆叠的瓶颈(BottleNeck)模块。该模块的架构根据模型的大小在宽度和深度上有所不同，由预定义的参数控制。C3 模块包含一个类似于瓶颈 CSP 的残差结构。它以两种方式处理输入特征图：通过双分支方法，其中一个分支使用两个标准卷积层，其他分支输出原始特征图，随后将输出连接起来，或者通过放弃残差路径并在标准卷积后直接输出特征图。C3 中的瓶颈模块以其强大的特征提取能力以及在解决梯度消失和梯度爆炸问题的能力而闻名。

C3Ghost 模块：标准卷积模块，通常包括一个规则的卷积层以及批处理归一化和激活函数，通常创建许多相似的特征映射，导致高计算需求和资源消耗。为了解决这个问题，GhostCovn 采用了两步方法。最初，它采用具有较小内核大小的标准卷积来生成具有较少通道的特征映射。随后，使用深度卷积(DepConv)来生成第一步未创建的附加特征映射。然后将这两个阶段的特征映射结合起来，产生最终的特征映射，类似于由标准卷积层产生的特征映射，但计算量和参数要少得多。

在我们提出的 C3Chost 模块中，GhostCovn 与 Ghost 瓶颈模块一起使用，其中包括 GhostCovn 和 DepConv 模块，并构成残差结构。GhostBottleNeck 有两个分支：第一个分支通过 1 × 1 步长为 1 卷积核的 GhostCovn 处理输入特征映射，然后将生成的特征映射添加到原始特征映射中。第二个分支引入了一个中间 DepConv 模块，其内核为 3 × 3，两个 GhostCovn 模块之间的步长为 2。残差路径遵循一个类似的 DepConv 模块，然后是一个步幅为 1 的标准 1 × 1 卷积。C3Chost 模块的整体结构类似于 C3 模块，但它用 GhostBottleNeck 代替了 BottleNeck。总体架构如图 3 所示。

C3CrossCovn 模块：虽然 GhostCovn 显著地简化了 C3 模块，但它可能导致沿信道方向的代表性信息丢失，从而影响模型精度。为了减轻这种情况，采用交叉卷积(CrossCovn)解决。CrossCovn 包括两个标准的卷积层，在特征映射上以交叉模式排列。它与传统的 k × k 滑动窗口卷积不同，第一层使用 1×k 核，其水平步长为 1，垂直步长为 s，第二层使用 k × 1 核，其两个维度的步长均为 s。CrossCovn 的示意图如图 4 所示。

为了评估标准卷积和交叉卷积在参数数量和计算量上的区别，我们建立了一个比较框架。对于这个分析，让我们考虑一个尺寸为 H×H×C 的正方形输入图像。卷积使用 k × k 的核大小，核的数量等于通道的数量 C，在图像边缘周围填充 p。公式 1 确定了通过标准卷积层进行处理所需的浮点计算需求和参数计数：

$$\begin{aligned} \mathrm{FLOPs}_1 &= k^2 C\left(\frac{W-k+2p}{s}+1\right)^2 \\ \mathrm{parameters}_1 &= k^2 C \end{aligned} \tag{1}$$

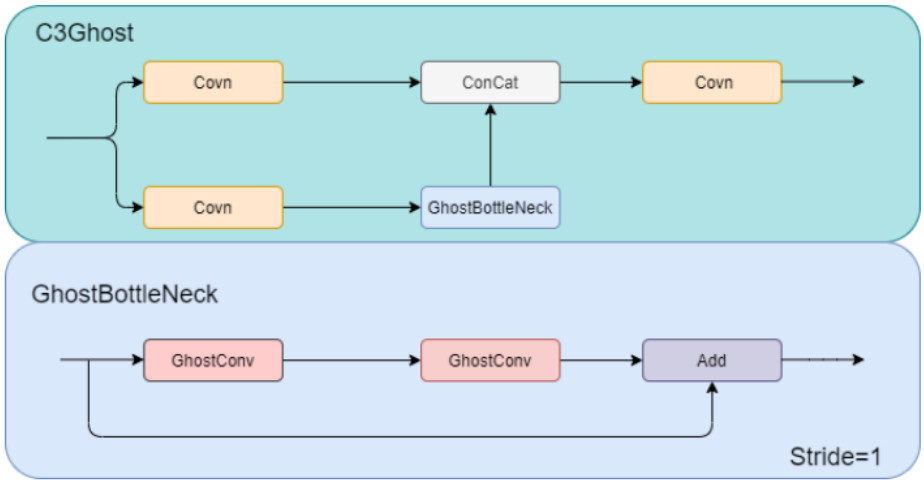

Figure 3. C3Ghost module
图 3. C3Ghost 模块

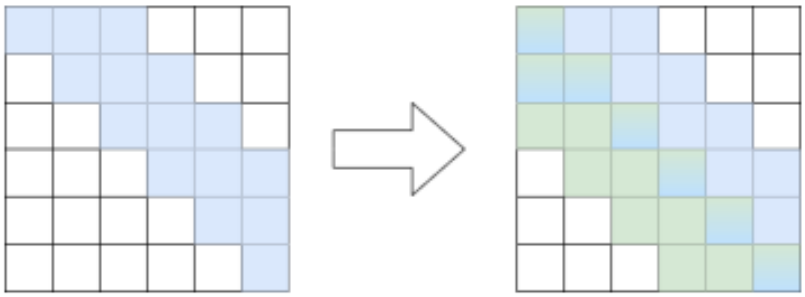

Figure 4. CrosssCovn module (stride = 1, kernel size = 3)
图 4. CrosssCovn 模块(步长为 1，内核大小为 3)

为了确定 CrossCovn 对图像进行单个操作所需的计算需求，我们做出如下假设：CrossCovn 包含 C 个双卷积核，第一个核的大小为 1 × k，第二个核的大小为 k × 1，步长为 s。公式 2 详细介绍了 CrossCovn 中计算负载的具体计算以及与这些设置相关的参数数量：

$$\begin{aligned} \text{FLOPs}_2 &= k^2 C\left(\frac{W-1+2p}{s}+1\right)\left(\frac{W-k+2p}{s}+1\right) \\ \text{parameters}_2 &= 2kC \end{aligned} \tag{2}$$

在 MS COCO 数据集中，所有图像都是 RGB，因此我们使用三个图像通道(C = 3)。为了确保足够大的接受域，我们将 k 设置为 3，s 设置为 1。在这些条件下，很明显，标准卷积需要更多的计算工作量，并且与 CrossCovn 相比，其参数数量约为 1.5 倍。尽管 CrossCovn 需要在单个特征映射上进行两次条纹核卷积操作，但它比标准卷积实现了更精细的特征提取和更丰富的特征信息。这种增强不仅提高了检测精度，而且显著降低了计算需求和参数数量，使其成为模型轻量化的最佳解决方案。C3CrossCovn 模块概述如图 5 所示。

3. 实验

3.1. 数据集和参数设置

本研究的实验采用 MS COCO 数据集。具体来说，使用了包含 118,287 张图像和 117,266 个标签的

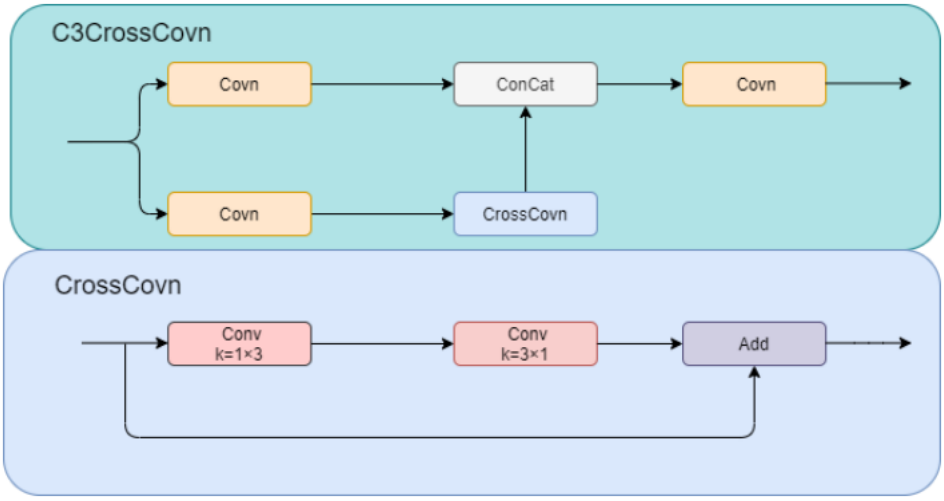

Figure 5. C3CrossCovn Module
图 5. C3CrossCovn 模块

COCO 训练集和包含 5000 张图像和 4952 个标签的 COCO 验证集。考虑到这些图像的不同大小，它们在实验中被统一调整为 640 × 640 像素。COCO 训练集包含了广泛的对象，共计 80 个对象类别，代表了日常生活中常见对象的全面集合。这使得 MS COCO 数据集在计算机视觉研究中具有广泛的应用价值。对于训练模型，使用 Adam 优化器，从学习率为 0.001 开始，增加到 0.01。为了提高参数更新的速度，我们将动量设置为 0.937。权重衰减对于训练中的正则化至关重要，它被小心地平衡在 0.0005 以避免过拟合或者模型拟合不足。训练方案包括最初的热身阶段，涵盖前 3 个阶段，然后是广泛的训练阶段，总共跨越 400 个阶段。该模型是用 Python 实现的，利用 PyTorch 框架，并在配备了四个 16GB V100 GPU 的云服务器上进行了培训。

3.2. 微小目标检测层的评价

通过在颈部网络中加入微小目标检测层，YOLOv5s-sm 在几个指标上优于基准 YOLOv5s，详见表 2。这种增强体现到 $Precision_{all}$ 增加 1%，$Recall_{all}$ 增加 2.2%，mAP@0.50 增加 1.8%，mAP@0.5:0.95 增加 1.4%。然而，这导致了更高的计算和参数要求，可能是由于颈部网络中额外的卷积模块创建了更大的特征映射。

Table 2. Detection results of tiny object detection layers
表 2. 小目标检测层检测结果

Models	$Precision_{all}$	$Recall_{all}$	mAP@0.50	mAP@0.5:0.95	F1	GFLOPs	Parameters (M)
YOLOv5s	67.7%	50.3%	55.7%	36.1%	0.577	16.6	7.23
YOLOv5s-sm	68.7%	52.5%	57.5%	37.5%	0.595	19.9	7.38

为了突出所提出的微小目标检测层的有效性，我们使用了来自 YOLOv5s 和 YOLOv5s-sm 模型的结果图像和头部网络热图。图 6 显示了(a)原始图像，(b) YOLOv5s 的推理结果，(c) YOLOv5s-sm 的推理结果表明 YOLOv5s-sm 在检测小目标时具有优越的能力。图 7 为来自 YOLOv5s 颈部网络的三个特征热图，分别对应大、中、小目标的检测层，图 8 为来自 YOLOv5s-sm 的对应层，包括微小目标检测层。值得注意的是，从图 8(c)和图 8(d)可以看出，YOLOv5s-sm 中的微小物体检测层对微小物体的关注更加细致。

(a) (b) (c)

Figure 6. Detection results of YOLOv5s and YOLOv5s-sm
图 6. YOLOv5s 和 YOLOv5s- sm 的检测结果

(a) (b) (c)

Figure 7. Heat map of YOLOv5s neck network outputs
图 7. YOLOv5s 颈部网络输出热图

(a) (b) (c) (d)

Figure 8. Heat map of YOLOv5s-sm neck network outputs
图 8. YOLOv5s-sm 颈部网络输出热图

3.3. 加入轻量级卷积模块的 YOLOv5s-QCK

为了说明轻量化策略对模型性能的影响，本节选取不同策略。这些不同修改的结果详列于表 3。

Table 3. Detection results of different improved models.
表 3. 不同改进模型的检测结果

Models	$Precision_{all}$	$Recall_{all}$	mAP@0.50	mAP@0.5:0.95	F1	GFLOPs	Parameters (M)
YOLOv5s	67.7%	50.3%	55.7%	36.1%	0.577	16.6	7.23
YOLOv5s-sm	68.7%	52.5%	57.5%	37.5%	0.595	19.9	7.38
YOLOv5s-QC	67.0%	53.7%	58.1%	38.0%	0.596	19.7	7.19
YOLOv5s-QK	70.4%	53.9%	59.6%	39.7%	0.611	25.5	9.70
YOLOv5s-QCK	71.2%	57.3%	60.3%	40.1%	0.635	25.3	9.49

与 YOLOv5s-sm 相比，YOLOv5s-QC 在 mAP@0.50 上提高了 0.6%，在 mAP@0.5:0.95 上提高了 0.5%，同时参数降低了 0.19 M。YOLOv5s-QK 比 YOLOv5s-sm 高 2.1% mAP@0.50 和 2.2% mAP@0.5:0.95。与

YOLOv5s-QK 相比，YOLOv5s-QCK 减少了 0.21 M 的参数和 0.2GFLOPs 的计算需求，同时在 mAP@0.50 上实现了 0.7%的改进，在 mAP@0.5:0.95 上实现了 0.4%的改进。这些结果表明，轻量化方法不仅降低了模型复杂度，而且提高了检测性能。虽然精度的提高通常会导致计算和参数需求的增加，但这里实现的轻量化策略有效地限制了这些增加。实验结果验证了所提出的每一个改进都对 YOLOv5s 模型的增强性能有积极的贡献，并证实了这些改进不是互斥的，而是互补的。

为了证明 YOLO-QCK 方法的有效性，我们使用从不同验证集中随机获取的两张测试图像将其与 YOLOv5s 进行比较，这些图像的特征是密集分布和均匀大小的对象。从图 9 可以看出，本文提出方法在检测框预测出来的目标值均优于基础的 YOLOv5s，尤其是在小目标框上的预测值对比，所以 YOLO-QCK 在检测小目标方面优于 YOLOv5s。

(a) YOLOv5s (b) YOLO-QCK

Figure 9. Visualized results of different detection models
图 9. 对比可视化结果

4. 结束语

本研究解决了目标检测领域的普遍挑战，并引入了一种新的方法 YOLO-QCK。与基准的 YOLOv5 相比，该方法显示出优越的检测性能，特别是在准确识别小物体方面。YOLO-QCK 通过在颈部网络中集成微小目标检测层，为了平衡增强检测和模型效率，采用轻量级策略，在骨干网中加入 C3CrossCovn 模块以降低模型复杂度。与基线 YOLOv5s 模型相比，我们新开发的模型 YOLO-QCK 在 MS COCO 验证数据集上显示出相当大的改进，mAP@0.5 增加了 4.6%，mAP@0.5:0.95 增加了 4%，同时保持模型尺寸紧凑，参数为 9.49 M。同时通过在模型上引入不同轻量化策略的实验结果对比证明这种策略不仅降低了复杂性，而且提高了准确性。接下来将继续优化该网络，通过研究注意力机制对结果提升的影响。

参考文献

[1] Zou, Z., Chen, K., Shi, Z., Guo, Y., and Ye, J. (2023) Object Detection in 20 Years: A Survey. *Proceedings of the IEEE*, **111**, 257-276. https://doi.org/10.1109/JPROC.2023.3238524

[2] Kaur, R. and Singh, S. (2023) A Comprehensive Review of Object Detection with Deep Learning. *Digital Signal Processing*, **132**, Article ID: 103812. https://doi.org/10.1016/j.dsp.2022.103812

[3] Xu, S., Zhang, M., Song, W., Mei, H., He, Q. and Liotta, A. (2023) A Systematic Review and Analysis of Deep Learning-Based underwater Object Detection. *Neurocomputing*, **527**, 204-232. https://doi.org/10.1016/j.neucom.2023.01.056

[4] Zhao, Q., Liu, B., Lyu, S., Wang, C. and Zhang, H. (2023) Tph-YOLOv5++: Boosting Object Detection on Drone-Captured Scenarios with Cross-Layer Asymmetric Transformer. *Remote Sensing*, **15**, Article 1687. https://doi.org/10.3390/rs15061687

[5] Mao, J., Shi, S., Wang, X. and Li, H. (2023) 3D Object Detection for Autonomous Driving: A Comprehensive Survey. *International Journal of Computer Vision*, **131**, 1-55. https://doi.org/10.1007/s11263-023-01790-1

[6] Zhang, L., Wang, G., Chen, M., Ren, F. and Shao, L. (2023) An Enhanced Noise-Tolerant Hashing for Drone Object Detection. *Pattern Recognition*, **143**, Article ID: 109762. https://doi.org/10.1016/j.patcog.2023.109762

[7] Jung, H.-K. and Choi, G.-S. (2022) Improved YOLOv5: Efficient Object Detection Using Drone Images under Various Conditions. *Applied Sciences*, **12**, Article 7255. https://doi.org/10.3390/app12147255

[8] Wozniak, M., Wieczorek, M., and Siłka, J. (2022) Deep Neural Network with Transfer Learning in Remote Object Detection from Drone. *Proceedings of the* 5*th International ACM Mobicom Workshop on Drone Assisted Wireless Communications for* 5*G and Beyond*, Sydney, 17 October 2022, 121-126. https://doi.org/10.1145/3555661.3560875

[9] 谷永立, 宗欣欣. 基于深度学习的目标检测研究综述[J]. 现代信息科技, 2022, 6(11): 76-81.

[10] 李雄飞, 王婧, 张小利, 等. 基于 SVM 和窗口梯度的多焦距图像融合方法[J]. 吉林大学学报(工学版), 2020, 50(1): 227-236.

[11] 刘国特, 伍伟权, 郭芳, 等. 基于改进级联 Gentle Adaboost 分类器的支柱绝缘子红外图像 AI 识别[J]. 高电压技术, 2022, 48(3): 1088-1095.

[12] Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014) Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 580-587. https://doi.org/10.1109/CVPR.2014.81

[13] Ren, S., He, K., Girshick, R. and Sun, J. (2015) Faster R-CNN: Towards Real-Time Object Detection with region Proposal Networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **39**, 1137-1149.

[14] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016) You Only Look Once: Unified, Real-Time Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 779-788. https://doi.org/10.1109/CVPR.2016.91

[15] Redmon, J. and Farhadi, A. (2018) YOLOv3: An Incremental Improvement. arXiv preprint arXiv: 1804.02767, 2018.

[16] Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B. and Belongie, S. (2017) Feature Pyramid Networks for Object Detection. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, July 21-26 2017, 936-944.

[17] Bochkovskiy, A., Wang, C.-Y. and Liao, H.-Y.M. (2020) YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint arXiv: 2004.10934, 2020.

[18] Liu, S., Qi, L., Qin, H., Shi, J. and Jia, J. (2018) Path Aggregation Network for Instance Segmentation. 2018 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 8759-8768. https://doi.org/10.1109/CVPR.2018.00913

[19] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., *et al*. (2020) An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv: 2010.11929, 2020.

[20] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020) End-to-End Object Detection with Transformers. Springer, Berlin, 213-229. https://doi.org/10.1007/978-3-030-58452-8_13

[21] Li, Y., Chen, Y., Wang, N. and Zhang, Z. (2019) Scale-Aware Trident Networks for Object Detection. *Proceedings of the IEEE*/*CVF International Conference on Computer Vision*, Seoul, 27 October-2 November 2019, 6054-6063.