文章编号:1007－6735(2025)03－0288－11　DOI: 10.13255/j.cnki.jusst.20240307001

基于 YOLOX 的轻量化目标检测算法及其应用

柴炜朕，　王朝立，　孙占全
（上海理工大学 光电信息与计算机工程学院，上海 200093）

摘要：目标检测算法被广泛应用于生产安全领域。针对现有的目标检测算法检测速度慢、在复杂施工环境下检测精度低的问题，提出一种改进的 YOLOX 检测算法。首先，基于轻量化卷积模组 Ghost moudle 重构主干网络，压缩模型参数量和计算量，提高检测速度；其次，在主干网络输出端嵌入坐标注意力机制，增强模型对于关键位置信息的学习能力；最后，在颈部网络中引入递归门控卷积，增强模型的空间位置感知能力，捕获图像中的长距离依赖关系。改进后的模型在数据集 Pascal VOC 和 SHWD 上进行实验验证，与基线模型相比，平均精度均值分别提升 1.69% 和 1.1%，模型参数量降低 18.8%，计算量降低 23.3%，帧率提升 7.6%。将本文模型部署在终端设备上，可应用于施工环境下的实时监控检测中。

关键词：目标检测；轻量化；坐标注意力；递归门控卷积
中图分类号：TP 391　　**文献标志码：**A

Lightweight object detection algorithm based on YOLOX and its application

CHAI Weizhen,　WANG Chaoli,　SUN Zhanquan
(*School of Optical−Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China*)

Abstract: Object detection algorithms are widely used in the field of production safety. To address the problems of slow detection speed and low detection accuracy in complex construction environments, an improved YOLOX detection algorithm was proposed. First, based on the lightweight convolution module Ghost moudle, the backbone network was reconstructed to compress the model parameters and computational complexity, thereby improving detection speed. Second, embedding the coordinate attention mechanism at the output of the backbone network to enhance the model's ability to learn key

收稿日期: 2024−03−07
基金项目: 国家自然科学基金资助项目 (62173232)；国防科工局基础研究项目 (JCKY2019413D001)
第一作者: 柴炜朕(1998—)，男，硕士研究生．研究方向：图像处理．E-mail：15995020039@163.com
通信作者: 王朝立(1965—)，男，教授．研究方向：人工智能、非线性控制．E-mail：clwang@usst.edu.cn

引文格式: 柴炜朕，王朝立，孙占全．基于 YOLOX 的轻量化目标检测算法及其应用[J]．上海理工大学学报，2025，47(3)：288-298.
Citation: CHAI Weizhen, WANG Chaoli, SUN Zhanquan．Lightweight object detection algorithm based on YOLOX and its application[J]．Journal of University of Shanghai for Science and Technology, 2025, 47(3)：288-298.

position information. Finally, recursive gated convolution was introduced into the neck network to enhance the model's spatial position perception ability and capture long-range dependencies in the image. The improved model are experimentally validated on the Pascal VOC and SHWD datasets, comparing with the baseline model, mean average precision increase by 1.69% and 1.1% respectively, model parameter count decrease by 18.8%, computational load decrease by 23.3%, and frame rate increase by 7.6%. Deploying the purposed model on terminal devices can be applied to real-time monitoring and detection in construction environments.

Keywords: *object detection*; *lightweight*; *coordinate attention*; *recursive gated convolution*

目标检测算法在生产安全领域应用广泛，它可以代替传统的人工监督方法，节约人力物力。在智慧工地建设中，将目标检测算法应用于工人安全帽佩戴检测，可以有效保障复杂施工环境下工人的生命安全。但在实际应用场景中，对算法的实时性和检测精度都有较高的要求。

传统的目标检测方法需要人工设计图像特征提取算子再进行检测，代表算法有 HOG[1] 和 DPM[2]，这种方法过程繁琐，模型通用性和鲁棒性较差，检测准确率较低。深度学习方法不用人为设置图像特征，可以根据图像本身学习其内在的属性，基于深度学习的目标检测算法已被广泛应用于生产安全领域，例如安全帽检测、人脸识别、火焰检测等。目前，基于深度学习的目标检测算法可大致分为 2 类：一类是两阶段（two-stage）目标检测算法，其代表有 Fast R-CNN[3]、Faster R-CNN[4]；另一类是单阶段（one-stage）的目标检测算法，其代表有 YOLO[5]、SSD[6] 和 ATSS[7] 等。两阶段检测算法的思路，是将可能存在物体的区域视为候选区域（region proposal），在此候选区域的基础上完成分类和回归任务。这种方法的网络结构复杂、计算量大，检测速度与单阶段检测算法差距很大，不能满足安全帽检测这一实时应用场景的需求。而在单阶段的检测算法中，省去了候选区域的提取工作，经过单阶段的网络直接完成分类和回归任务，因此其检测速度大大提高，可以满足实时检测应用场景的需求。在单阶段检测算法中，YOLO 系列检测方法既可以保持较高的精度，又有很快的检测速度，因此在安全帽检测中应用最为广泛。张业宝等[8] 在 SSD 的基础上，采用 VGG16[9] 预训练模型进行迁移学习，加速训练过程，提高了检测速度和检测精度。谢国波等[10] 在 YOLOv4[11] 的基础上，融合通道注意力机制、增加特征尺度，改善了算法的检测精度，但检测速度和精度还显不足，主要原因是 YOLOv4 使用的是 DarkNet-53 主干网络。吕宗喆等[12] 在 YOLOv5 的基础上，对边界框损失函数和置信度损失函数进行优化，改善了对密集小目标特征的学习效果。

YOLOv3[13] 是 YOLO 系列的集大成者，但其需要人工设计检测框的大小。在 FCOS[14] 中，作者将特征图像素坐标映射回到原图，通过 1 个解耦头直接完成分类和回归任务，省去了复杂的锚框设计。Ge 等[15] 借鉴了 FCOS 的解耦头设计，在 YOLOv3 的基础上改进得到 YOLOX。该算法引入 SimOTA 标签分配策略筛选正样本、加快网络推理速度，采用无锚框检测思路省去了复杂的锚框设计、节省了计算资源，并且采用解耦头分别预测回归任务和分类任务，提高了检测精度。YOLOX 在保证高检测精度的前提下，检测速度快、模型参数量小、部署简便，非常适用于实际场景中的实时检测。本文在 YOLOX 的基础上，对算法继续进行改进，进一步提升网络的检测精度和检测速度，主要工作为:

a. 在主干网络中引入了轻量化卷积模组 Ghost moudle[16]，构建了新的主干网络 G-Backbone，在保证检测精度几乎没有损失的前提下，大大压缩了模型参数量和计算量，提高了检测速度。

b. 在主干网络的输出端融入了坐标注意力（coordinate attention，CA）机制[17]，有效解决了安全帽检测中复杂背景带来的特征提取困难的问题，减少了关键特征信息的损失，同时还能学习到空间位置信息，有助于模型快速定位与识别。

c. 提出了基于递归门控卷积模组 g^nConv[18] 的特征提取模块 g-CSP，将其融入到颈部网络中构建了新的颈部网络 g-Neck，增强了网络的空间位置

感知能力，改善了多目标的检测，且在 g^{n}Conv 中还采用了 7×7 的大卷积核，有利于扩大网络的感受野，捕获长距离依赖关系。

1 YOLOX 目标检测算法

YOLOX 的网络结构如图 1 所示，其主体结构由主干网络(backbone)、颈部网络(neck)、检测头(head)3 部分构成。在输入端，使用 Mosaic 和 Mixup 数据增强策略提高数据的丰富性，缓解因数据不足而造成的网络过拟合现象；主干网络为 Darknet-53，用于初步提取图像特征，主体结构为 Focus 和 CSP 组件。Focus 的操作过程如图 2 所示，具体为在输入图片上隔像素取值，以进行下采样，采样得到的特征图通道数变为原来的 4 倍，最后再经过 1 个 3×3 卷积提取特征。Focus 操作可以把原来单通道上的信息拓展到通道空间上，并且减少了下采样过程中的信息丢失。CSP 组件为融合了残差块的全卷积特征提取模块，在主干网络中，通过多次堆叠特征提取模块，可以加深网络层数，扩大网络感受野，提取深层次的图像信息。输入图像经过主干网络后，分别输出 3 个不同尺度的特征图，随后进入路径聚合特征金字塔网络(path aggregation-feature pyramid network，PA-FPN)中进行多尺度的特征融合。首先，在特征金字塔网络(feature pyramid network，FPN)中采用自顶向下的方式，将深层信息上采样后与浅层信息融合；其次，在路径聚合网络(path aggregation network，PAN)中则采用自底向上的方式，将浅层信息下采样后与深层信息融合，融合过后依然得到 3 个不同尺度的特征图，这 3 个尺度的特征图分别对应着不同大小的感受野并用于检测不同大小的物体；最终，将颈部网络输出的特征再输入到图 3 所示的解耦头中，在解耦头中将特征信息解耦，分为两条支路，分别负责预测分类任务和回归任务，预测结果包含分类分数、位置参数和置信度 3 部分。

2 模型改进

2.1 轻量化主干网络 G-backbone

常规卷积如图 4 所示，对于输入数据 $X \in R^{c\times h\times w}$，$c$、$h$、$w$ 分别表示输入特征图的通道数、高和宽，常规卷积操作如下：

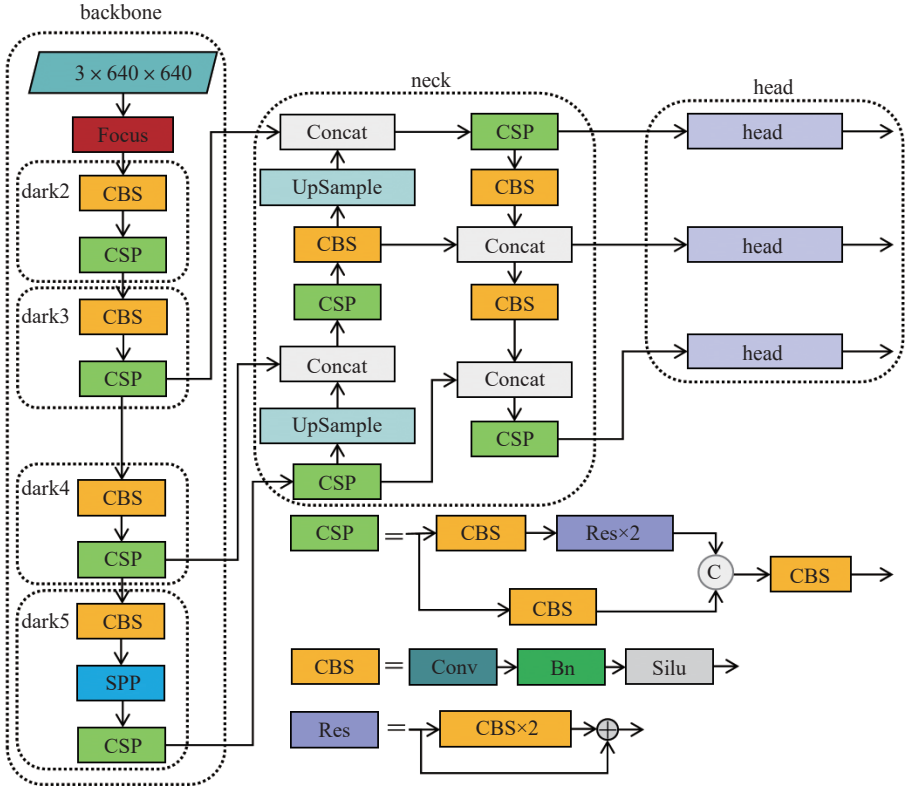


图 1 YOLOX 网络模型

Fig.1 YOLOX network

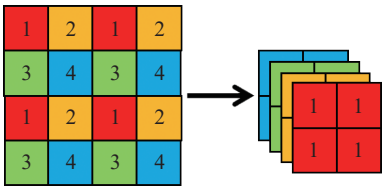


图 2　Focus 操作

Fig.2　The operation of Focus

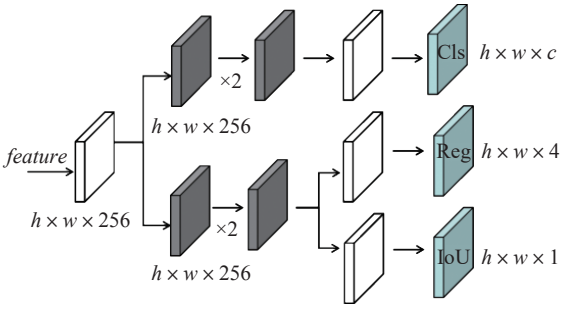


图 3　解耦头

Fig.3　Decouped head

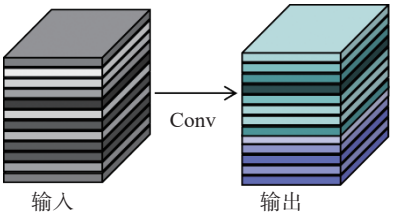


图 4　常规卷积

Fig.4　Traditional convolution

$$Y = X * f + b \tag{1}$$

式中：$*$表示卷积操作；b 为偏置项；$Y \in R^{n' \times h' \times w'}$ 为输出特征图，n'、h'、w'分别表示输出特征图的通道数、高和宽；$f \in R^{c \times k \times k \times n}$ 为卷积操作，$k \times k$ 为卷积核大小，n 为卷积核个数。

常规卷积操作的输出特征图中往往包含很多相似的冗余信息，为压缩模型参数量，提高网络的检测速度，本文引入了图 5 所示的轻量化卷积 Ghost moudle 重建主干网络。Ghost moudle 先通过常规卷积操作，输出 m 层初始特征图 $Y' \in R^{m \times h' \times w'}$，如式(2)所示：

$$Y' = X * f' \tag{2}$$

式中，$f' \in R^{c \times k \times k \times m}$ 为卷积操作且 $m \leqslant n$。再将其通过一系列廉价的线性变换操作，得到相似的 $(n-m)$ 层特征图，如式(3)所示：

$$y_{ij} = \Phi_{i,j}(y_i'), \quad \forall i = 1, \cdots, m \quad j = 1, \cdots, s \tag{3}$$

式中：y代表通过线性变换得到的$(n-m)$层相似特征图，$s = n/m$为扩充的倍数；Φ为线性变换，其具体实现为卷积核大小为$d \times d$的深度可分离卷积。最后，把初始特征图Y'恒等映射后的结果和线性变换后的结果y拼接作为最终的输出特征图。

$$K_{\text{conv}} = cnk^2 h' w' \tag{4}$$

考虑到 $d \approx k$以及 $s \ll c$，常规卷积与 Ghost moudle 的计算量之比为

$$\begin{aligned} r_{\text{s}} &= \frac{nh'w'ck^2}{\frac{n}{s}h'w'ck^2 + (s-1)\frac{n}{s}h'w'd^2} \\ &= \frac{ck^2}{\frac{1}{s}ck^2 + \frac{s-1}{s}d^2} \approx \frac{sc}{s+c-1} \approx s \end{aligned} \tag{5}$$

由此可见，常规卷积的计算量是 Ghost moudle 的 s 倍。因此，通过 Ghost moudle 重新构建主干网络，可以达到减少计算量、加速计算的目的，本文中 s 取 2。

基于 Ghost module 构建的特征提取模块 Ghost bottleneck 的结构如图 6 所示。当步长 stride=1 时，Ghost bottleneck 由 2 个 Ghost moudle 构成，输入输出维度相同。当 stride=2 时，Ghost bottleneck 由 2 个 Ghost moudle 和 1 个深度可分离卷积块组成，残差连接部分加入卷积操作降维，在第一个 Ghost moudle 后，同样采用深度可分离卷积降维。此时，模块除了可以进行特征提取，还兼具下采样的功能。基于 Ghost bottleneck重构的主干网络结

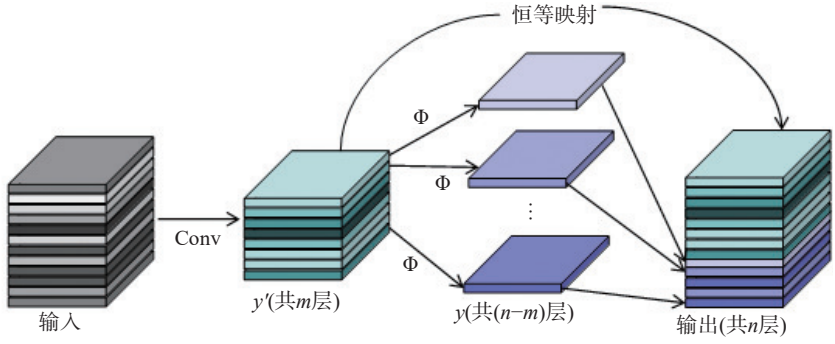


图 5　轻量化卷积 Ghost moudle

Fig.5　Lightweight convolutional Ghost moudle

构如表 1 所示，将原网络的 CBS和 CSP 模块替换为更加轻量的 Ghost bottleneck，考虑到轻量化网络可能会带来信息的丢失问题，在主干网络中嵌入 SE(squeeze-and-excitation)[19] 模块，增强关键特征，提高主干网络的特征提取能力。更改后，主干网络的模型参数量大幅下降，检测速度更快。

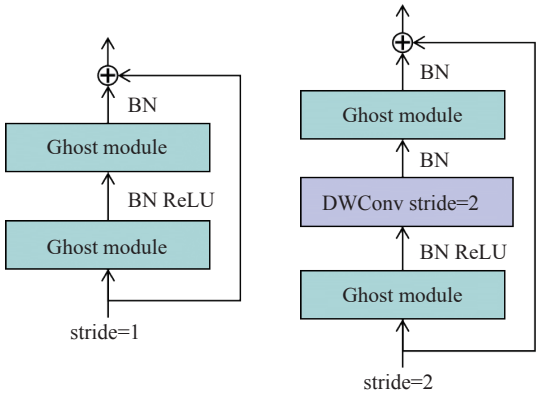


图 6 Ghost bottleneck 结构

Fig.6 The structure of Ghost bottleneck

表 1 重构的主干网络

Tab.1 Reconstructed backbone network

输入	运算操作	步长
$640^2\times3$	Focus	2
$320^2\times32$	CBS	2
$160^2\times64$	Ghost bottleneck	1
$160^2\times64$	Ghost bottleneck	2
$80^2\times128$	Ghost bottleneck	1
$80^2\times128$	SE	1
$80^2\times128$	Ghost bottleneck	2
$40^2\times256$	Ghost bottleneck	1
$80^2\times128$	SE	1
$40^2\times256$	Ghost bottleneck	2
$20^2\times256$	SPP	1
$20^2\times256$	Ghost bottleneck	1
$20^2\times256$	SE	1

2.2 坐标注意力机制

为避免安全帽检测中复杂背景带来的冗余信息感干扰，弥补 YOLOX 网络检测精度的不足，使网络关注到更有意义的信息，在主干网络的输出端引入注意力机制。传统的通道注意力机制 SE 采用了全局池化对空间信息进行编码，但它将全局信息压缩到通道描述中，因此很难保存位置信息，而位置信息在视觉任务中对于捕获空间位置结构至关重要。在本文中，引入了 CA 机制，它采用“分而治之”的思想，将全局池化分解为垂直方向和水平方向上的平均池化，进而实现位置信息的编码，其结构如图 7 所示。

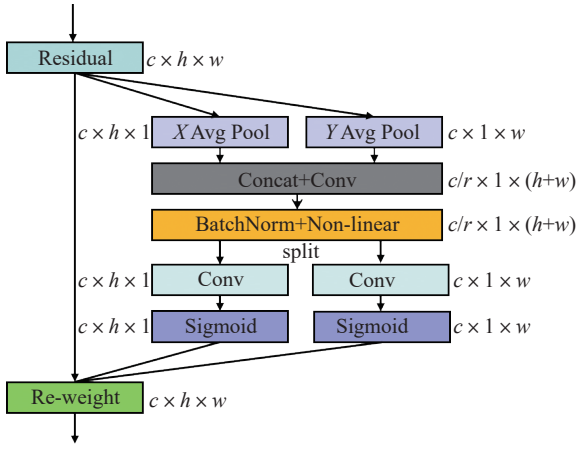


图 7 坐标注意力机制

Fig.7 Coordinate attention mechanism

对于输入数据 $X\in R^{c\times h\times w}$，分别使用平均池化核 $(H,\ 1)$ 和 $(1,\ W)$ 编码水平方向和垂直方向的信息，则第 c 个通道在高 h 处的编码信息为

$$z_c^h(h)=\frac{1}{W}\sum_{0\leqslant i\leqslant W}x_c(h,j) \tag{6}$$

式中：$x_c(h,j)$指输入特征图第 c 个通道上位置坐标 (h,j)处的像素值。

同样地，第 c 个通道在宽 w 处的编码信息为

$$z_c^w(w)=\frac{1}{H}\sum_{0\leqslant j\leqslant H}x_c(j,w) \tag{7}$$

式中：$x_c(j,w)$指输入特征图第 c 个通道上位置坐标 (j,w)处的像素值。

随后，将两部分编码信息拼接，经过 1 个 1×1卷积层F_1和非线性激活函数层δ，将特征图通道数压缩为c/r，生成包含位置信息的中间特征层 $f\in R^{c/r\times(h+w)}$，即

$$f=\delta\left\{F_1([z^h,z^w])\right\} \tag{8}$$

再将f沿空间维度分解为 2 个向量$f^h\in R^{c/r\times h}$和 $f^w\in R^{c/r\times w}$，后再经过 2 个1×1卷积层 F_h、F_w和 2 个 Sigmoid 激活函数层δ，将通道数还原为c，得到 2 个独立空间方向的注意力权重为

$$\begin{aligned}g^h&=\sigma\left[F_h(f^h)\right]\\g^w&=\sigma\left[F_w(f^w)\right]\end{aligned} \tag{9}$$

最终的输出为

$$y_c(i,j)=x_c(i,j)\times g_c^h(i)\times g_c^w(j) \tag{10}$$

区别于通道注意力模块仅考虑不同通道的重要程度，CA 模块考虑了空间信息的编码，从而可以让模型更加准确地定位感兴趣对象的确切位置。本文中，将 CA 模块引入到图 1 中 dark3、dark4、dark5 的输出端，可以显著提升模型的检测精度。

2.3　融合空间信息的颈部网络 g-Neck

YOLOX 颈部网络用于进一步的特征提取与特征融合，此过程中的卷积操作全部为标准的卷积操作，但标准的卷积操作没有考虑到空间之间的相互作用，在经过多次的卷积、上采样、拼接操作后，会造成细节空间信息损失，不利于对多目标和小目标物体的检测。为解决以上问题，本文引入了递归门控卷积模组 g^{n}Conv 模块，设计了新的 CSP 组件和新的颈部网络 g-Neck。

g^{n}Conv 是 1 个实现长期和高阶空间交互作用的模块，n 代表空间交互的阶数。此模块具有类似于自注意力 self-attention[20] 的作用，它借鉴了 Vit[21] 中对于空间交互作用的建模。Vit 的成功主要归因于自注意力机制对视觉数据中的空间交互作用进行了适当的建模。但是，自注意力操作引入的巨大参数量，对于要求实时性的 YOLOX 模型来说并不实用。而 g^{n}Conv 通过标准卷积、线性映射和矩阵元素相乘等廉价操作实现空间交互，计算量代价更小。本文引入的三阶 g^{3}Conv 的结构如图 8 所示。

g^{n}Conv 的基础结构为 $n=1$ 的一阶门控卷积 gConv，对于输入特征图 $X\in R^{c\times h\times w}$，gConv 的输出 y 为

$$[p_0^{hw\times c},q_0^{hw\times c}]=\Phi_{\text{in}}(x)\in R^{hw\times 2c}$$
$$y=\Phi_{\text{out}}[f(q_0)\otimes p_0]\in R^{hw\times c} \tag{11}$$

式中：Φ_{in} 和 Φ_{out} 为执行通道混合操作的线性映射操作，具体实现是卷积核大小为 1×1 的常规卷积；f 为卷积核大小为 7×7 的深度可分离卷积，大卷积核旨在增大网络的感受野，更容易捕获长期依赖关系；p_0 和 q_0 是将线性映射操作的结果在通道维度上切分后的特征层；$\otimes$ 为按元素相乘的操作。以此类推，将此过程重复 n 次，即可得到 n 阶的递归门控卷积 g^{n}Conv。图 8 中演示的为三阶递归门控卷积。通过这种建模通道间信息交互的做法，可以极大地提升模型的空间建模能力。

本文基于由 g^{n}Conv 卷积模块构建的特征提取模块 g-Block，重新设计了新的特征提取组件 g-CSP。g-Block 模块借鉴了 self-attention 的元架构设计，其结构如图 9 所示，它包含了 1 个空间混合模块 HorBlock 和 1 个前馈网络（feed-forward network，FFN）。空间混合模块 HorBlock 是 1 个归一化层和 g^{n}Conv 组成的残差块，完成空间交互建模；FFN 为 1 个归一化层和多层感知机（multilayer perceptron，MLP）构成的残差块。g-Block 旨在用简单的操作建模长距离的空间信息，使其适用于轻量化的网络中。

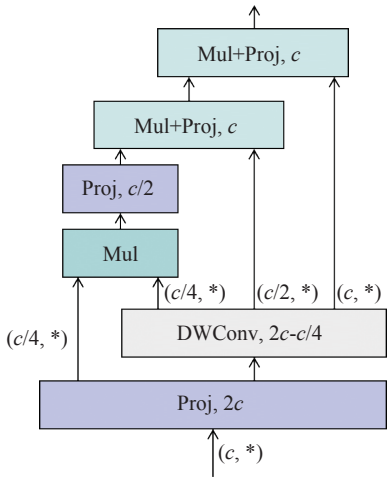


图 8　递归门控卷积模组

Fig.8　Recursive gated convolution moudle

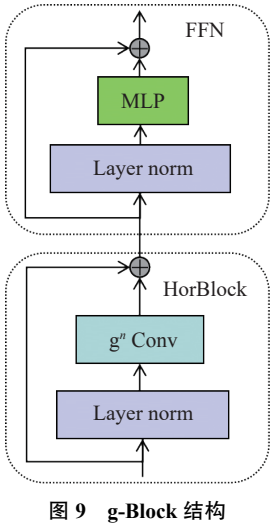


图 9　g-Block 结构

Fig.9　The structure of g-Block

重新设计的特征提取组件 g-CSP 结构如图 10 所示。在 g-CSP 中，输入特征先经过 1 个常规卷积块提取特征，后进入 g-Block 和深度可分离卷积块中提取特征，将其与另一条支路的常规卷积块提取的特征拼接后，再次输入到 1 个深度可分离卷积块中并输出最终结果。引入深度可分离卷积

的目的是降低参数量。本文基于 g-CSP 组件，设计了新的颈部特征融合网络 g-Neck，其结构如图 11 所示，C3、C4、C5 分别为从主干网络输出的 3 个尺度的特征层，其尺寸大小分别为 80×80×128、40×40×256、20×20×512。3 个尺度的特征经过上采样、下采样和拼接融合后，输出 3 个特征层 P3、P4、P5，其尺寸大小分别为 80×80×128、40×40×256、20×20×512。图 1 所示的 YOLOX 颈部网络采用了 CSP 组件进行特征提取与融合。本文将颈部网络中原始的 CSP 组件替换为重新设计的 g-CSP 组件，旨在改善模型的空间特征提取能力。

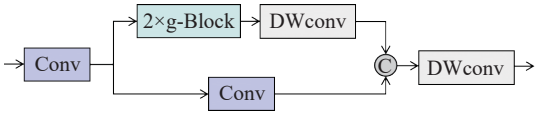


图 10 g-CSP 结构

Fig.10 The structure of g-CSP

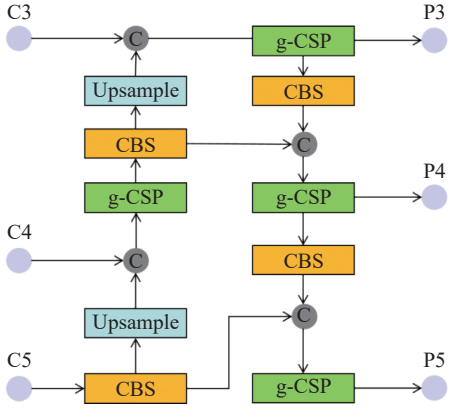


图 11 g-Neck 结构

Fig.11 The structure of g-Neck

3 实验结果与分析

3.1 数据集与实验环境

实验数据采用 2 个开源的数据集：第一个数据集为 SHWD 安全帽数据集，针对生产安全领域的安全帽检测问题；第二个数据集为 PASCAL VOC07+12[22-23] 通用数据集，用于对改进后的算法做进一步的验证。其中 SHWD 安全帽数据集包含 7581 张图片，训练集和测试集按 7∶3 划分，训练集包含 5 307 张图片，测试集包含 2274 张图片，包含 hat 和 person2 个类别；Pascal VOC 数据集由 Pascal VOC2007 和 Pascal VOC2012 这 2 个数据集合并得到，训练集由 2 个数据集共同组成，一共含 16 551 张图片，测试集由 Pascal VOC2007 提供，包含 4952 张图片，包含 cat、person、dog 等 20 个类别。

实验在 Linux 操作系统上进行训练和测试，处理器型号为 Intel(R) Xeon(R) Platinum 6164 CPU @ 1.90 GHz，内存大小为 24 GB，显卡型号为 NVIDIA GeForce RTX 3090，深度学习框架为 Pytorch，每个模型从头开始训练，网络输入大小为 640×640，epoch 为 200 轮，初始学习率为 0.002 5。在前 5 轮训练中，冻结学习率，使网络熟悉数据；后 195 轮训练中优化器设置为 SGD，其动量和衰减系数分别为 0.9 和 0.000 5，单次训练样本数量 batch size 为 16。训练数据采用 Mosaic 和 Mixup 数据增强以增加数据的多样性：Mosaic 数据增强为随机选择 4 张图片，将其缩放后拼接为 1 张图片；Mixup 数据增强为随机选择 2 张图片，将其按一定比例融合为 1 张图片。

3.2 评价指标

为了衡量模型检测效果，本文采用模型参数量衡量模型大小，采用每秒 10 亿次的浮点运算数表示模型的计算量，采用帧率衡量模型的检测速度，采用平均精度 A 和平均精度均值 M_{AP} 衡量模型的检测精度。A 与 M_{AP} 综合考量了精准率 P 和召回率 R：P 是指正确预测某类的数量占全部预测为该类的数量的比例；R 是指正确预测某类的数量占实际上全部为该类的数量的比例。A 可以用来评估每一类的检测效果，它是以召回率为横轴、以精确率为纵轴组成的曲线围成的面积。将所有类的 A 值取均值可以得到 M_{AP}。计算公式为

$$\begin{cases} P = \dfrac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}}} \\ R = \dfrac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}} \\ A = \displaystyle\int_0^1 P(R)\mathrm{d}R \\ M_{\mathrm{AP}} = \dfrac{1}{a}\displaystyle\sum_{i=1}^{a} A_i \end{cases} \tag{12}$$

式中：N_{TP} 为正确预测为正类的样本数量；N_{FP} 为错误预测为正类的样本数量；N_{FN} 为漏检的正类样本数量；a 为所有类别数。

YOLOX 的损失函数包含 3 部分，分别为分类损失、置信度损失和锚框的回归损失，计算公式为

$$L = \frac{1}{N_{\mathrm{pos}}} L_{\mathrm{CE}}(Q_{\mathrm{Cls}}^{\mathrm{pos}}, G_{\mathrm{Cls}}^{\mathrm{pos}}) + \frac{\lambda_1}{N_{\mathrm{pos}}} L_{\mathrm{CE}}(Q_{\mathrm{Obj}}^{\mathrm{pos+neg}}, G_{\mathrm{Obj}}^{\mathrm{pos+neg}}) + \frac{\lambda_2}{N_{\mathrm{pos}}} L_{\mathrm{CIoU}}(Q_{\mathrm{Reg}}^{\mathrm{pos}}, G_{\mathrm{Reg}}^{\mathrm{pos}}) \tag{13}$$

式中：L_{CE} 和 L_{CIoU} 分别表示交叉熵损失函数和 CIoU[24] 损失函数；N_{pos} 表示正样本的数量；Q 和 G 分别表示预测样本和标签；上标 pos 和 neg 分别表示训练过程中的正样本和负样本；λ_1 和 λ_2 分别表示置信度损失和锚框回归损失的权重值；下标 Cls 表示分类任务，Obj 表示置信度预测任务，Reg 表示锚框回归任务。图 12 演示了预测框 B 与目标框 B^{gt} 的相交情形，ρ 为两框中心点的距离，l 为两框的最小外接矩形的对角线的长度，w^{gt} 为目标框 B^{gt} 的宽，h^{gt} 为目标框 B^{gt} 的高。CIoU 损失函数计算式为

$$\begin{cases} L_{CIoU} = 1 - I + \dfrac{\rho^2(u,u^{gt})}{l^2} + \alpha v \\ \alpha = \dfrac{v}{(1-I)+v} \\ v = \dfrac{4}{\pi^2}\left(\arctan\dfrac{w^{gt}}{h^{gt}} - \arctan\dfrac{w}{h}\right)^2 \end{cases} \tag{14}$$

式中：I 为交并比，即 2 个矩形框交集的面积与并集的面积之比；u、u^{gt} 分别表示预测框和目标框的中心点；α 为权重系数；v 为高宽比的一致性。图 13 展示了训练过程中的损失变化曲线，total_loss 代表总损失，ciou_loss 代表锚框回归损失，conf_loss 代表置信度损失，cls_loss 代表分类损失。可以发现，训练到 200 轮 epoch 时，4 项损失值均已收敛。

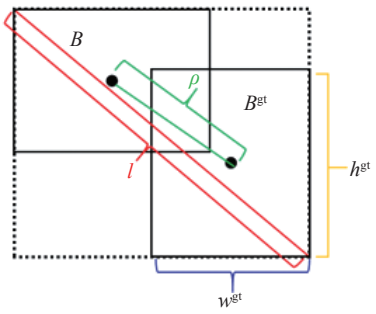


图 12　两框相交情形

Fig.12　Intersection of two boxes

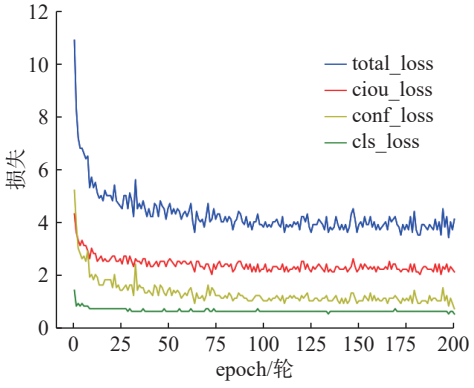


图 13　损失变化曲线

Fig.13　The curves of loss change

3.3　消融实验

为检验模型的有效性，在训练参数相同的情况下，对 3 处改进进行消融实验。在 SHWD 数据集上的结果如表 2 所示，A_1 和 A_2 分别表示对佩戴安全帽和未佩戴安全帽的工人的检测精度。实验 1 为本文基线模型 YOLOX 的实验结果；实验 2 采用轻量化的特征提取模块重新构建了主干网络，改进模型的参数量下降了 21.5%，计算量下降了 24.5%，检测速度提高了 14.1%，检测精度基本保持不变；实验 3 添加了 CA 模块，较基线网络的平均精度均值提升了 0.9%，说明 CA 模块可以让模型更加关注网络感兴趣的区域、过滤冗余的信息、提高模型的性能，而添加 CA 后的模型参数量和计算量基本保持不变；实验 4 采用空间交互能力更强的 g-Neck 颈部特征融合网络替换原网络，较基线网络的平均精度均值提高了 0.82%，其原因在于模型学习了更多的空间位置信息，定位能力更强，这对于目标检测任务而言至关重要，虽然模型参数量和计算量稍稍增加，牺牲了少量的检测速度，但仍满足使用需求；实验 5 替换轻量化主干网络的同时，添加了坐标注意力 CA 模块，结

表 2　SHWD 数据集上的结果

Tab.2　The results on SHWD dataset

实验	G-backbone	CA	g-Neck	参数量/MB	计算量/G	帧率/(f·s^{-1})	A_1/%	A_2/%	M_{AP}/%
1				8.94	28.48	92	86.15	86.49	86.32
2	√			7.02	21.50	105	86.14	86.42	86.28
3		√		8.97	28.50	89	87.01	87.43	87.22
4			√	9.15	28.79	87	86.96	87.32	87.14
5	√	√		7.05	21.52	102	87.02	87.38	87.20
6(本文)	√	√	√	7.26	21.83	99	87.97	88.05	88.01

果显示，CA 模块在只增加少量参数量和计算量的情况下，弥补了轻量化网络带来的检测精度损失；实验 6 为本文最终改进模型，同时添加了 3 处改进，较基线网络的参数量下降 18.8%，计算量下降 23.3%，检测速度提高 7.6%，平均精度均值提高 1.69%，证明最终改进模型是一种更快、更强的检测模型。图 14 为平均精度均值变化曲线，蓝色代表基线模型 YOLOX，红色代表本文最终改进模型。结果表明，改进后的模型以更高的精度收敛。

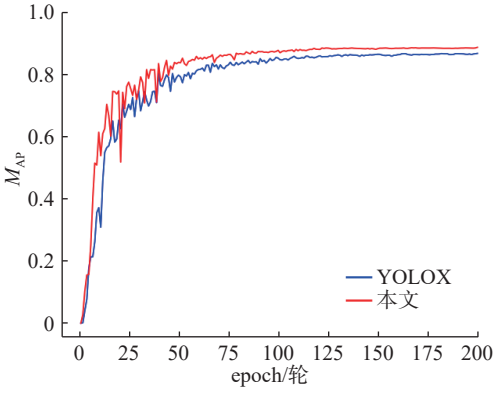


图 14 平均精度均值变化曲线

Fig.14 The curves of mean average precision change

在基线模型上分别添加 3 种不同的注意力机制 SE、CBAM[25]、CA，实验结果如表 3 所示。SE 采用全局池化的方法将通道信息压缩，没有考虑到位置信息；CBAM 在 SE 的基础上尝试通过卷积操作捕获空间位置信息，但卷积操作在建模长期依赖关系时效果不佳；CA 选择在特征图的 h 和 w 方向上压缩位置信息，再将其嵌入到通道注意力中，效果更佳。实验结果表明，添加 CA 相较于添加 SE 和 CBAM，平均精度均值分别提高了 0.28% 和 0.24%。

表 3 添加不同注意力机制的结果对比

Tab.3 Comparison of results with different attention mechanism

实验	A_1/%	A_2/%	M_{AP}/%
+SE	86.73	86.91	86.94
+CBAM	86.77	86.97	86.98
+CA	87.01	87.43	87.22

3.4 对比实验

为进一步验证模型的有效性，本文在公开数据集 SHWD 和 Pascal VOC 上与现阶段目标检测的主流模型进行性能比较，选用的对比模型有两阶段检测算法 Fast R-CNN、Faster R-CNN 和单阶段检测算法 SSD300、YOLOV3、YOLOV4、YOLOV5、YOLOX 这 7 个主流模型，对比结果如表 4 和表 5 所示。在数据集 SHWD 上，本文改进模型的平均精度均值可以达到 88.0%，检测速度可以达到 99 f/s，2 项指标在所有模型中均为最高。在 Pascal VOC 数据集上，本文改进模型的平均精度均值达到 83.0%，相较于基线网络提高 1.1%，检测速度为 99 f/s，2 项指标同样在所有模型中最高。

表 4 不同模型在 SHWD 数据集上的结果对比

Tab.4 Comparison of results for different models on SHWD dataset

方法	M_{AP}/%	帧率/(f·s^{-1})
Faster R-CNN	72.8	9
SSD300	74.3	59
YOLOv3	79.3	45
YOLOv4	80.5	73
YOLOv5	84.2	90
YOLOX	85.9	92
本文	88.0	99

表 5 不同模型在 Pascal VOC 数据集上的结果对比

Tab.5 Comparison of results for different models on Pascal VOC dataset

方法	M_{AP}/%	帧率/(f·s^{-1})
Fast R-CNN	70.0	1
Faster R-CNN	73.2	9
SSD300	74.3	59
YOLOv3	76.5	45
YOLOv4	77.9	73
YOLOv5	80.6	90
YOLOX	81.9	92
本文	83.0	99

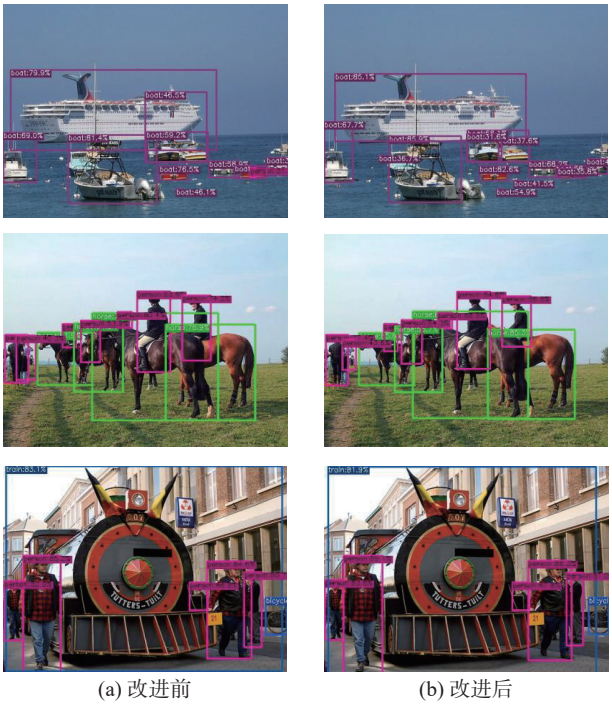
在 SHWD 数据集和 Pascal VOC 数据集中，分别选取 3 组图片，分别使用改进后的 YOLOX 算法和原始 YOLOX 算法进行检测。在 SHWD 数据集上的对比结果如图 15 所示，第 1 组和第 2 组图片中：原算法存在被遮挡物的漏检情况；改进后的算法检测出了被遮挡的安全帽，说明改进后的算法特征提取能力更强。第 3 组图片中：原算法不

仅存在远处小目标的漏检，还存在错检的情况；改进后的算法不仅检测到了图片较远处的小目标物体，错检率也明显下降。与原算法相比，改进后的算法检测置信度得分明显提高，检测更加精确。在 Pascal VOC 数据集上的对比结果如图 16 所示，第 1 组图片中，改进后的算法检测置信度较高，且错检的情况减少；第 2 组和第 3 组图片检测结果相同，但是改进后的算法检测置信度同样提高。

(a) 改进前　　(b) 改进后

图 15　SHWD 数据集可视化结果对比

Fig.15　Comparison of visualization results on SHWD dataset



(a) 改进前　　(b) 改进后

图 16　Pascal VOC 数据集可视化结果对比

Fig.16　Comparison of visualization results on Pascal VOC dataset

4　结 论

为提高实时检测算法的检测速度、检测精度，并将其应用于生产安全领域，本文在基线模型 YOLOX 的基础上提出 3 点改进。首先，将主干网络替换为更轻量、速度更快的 G-backbone，加快检测速度；其次，在主干网络的输出端引入 CA，使得模型特征提取能力增强，可以更好地关注到网络感兴趣的区域，过滤冗余信息；最后，在颈部网络中融合了空间建模能力更强的 g-CSP，使得模型更容易学习到空间位置信息，捕获图像中的长距离依赖关系。在 SHWD 和 Pascal VOC 数据集上分别进行实验验证，结果表明 3 种改进策略均起到效果，最终的改进模型相较于基线网络，不仅检测精度有提高，模型参数量和计算量也有所减少，而且检测速度更快，可以应用于施工环境下的实时监控检测中。

参考文献：

[1] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). San Diego: IEEE, 2005: 886–893.

[2] FELZENSZWALB P, MCALLESTER D, RAMANAN D. A discriminatively trained, multiscale, deformable part model[C]//2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage: IEEE, 2008: 1–8.

[3] GIRSHICK R. Fast R-CNN[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 1440–1448.

[4] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137−1149.

[5] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 779–788.

[6] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector[C]//14th European Conference on Computer Vision. Amsterdam: Springer, 2016: 21–37.

[7] ZHANG S F, CHI C, YAO Y Q, et al. Bridging the gap

between anchor-based and anchor-free detection via adaptive training sample selection[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 9756–9765.

[8] 张业宝, 徐晓龙. 基于改进 SSD 的安全帽佩戴检测方法[J]. 电子测量技术, 2020, 43(19): 80−84.

[9] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[DB/OL]. [2015-04-10]. https://arxiv.org/abs/1409.1556.

[10] 谢国波, 唐晶晶, 林志毅, 等. 复杂场景下的改进 YOLOv4 安全帽检测算法 [J]. 激光与光电子学进展, 2023, 60(12): 139−147.

[11] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal speed and accuracy of object detection[DB/OL]. [2020-04-23]. https://arxiv.org/abs/2004.10934.

[12] 吕宗喆, 徐慧, 杨骁, 等. 面向小目标的 YOLOv5 安全帽检测算法 [J]. 计算机应用, 2023, 43(6): 1943−1949.

[13] REDMON J, FARHADI A. YOLOv3: An incremental improvement[DB/OL]. [2018-04-08]. https://arxiv.org/abs/1804.02767.

[14] TIAN Z, SHEN C H, CHEN H, et al. FCOS: fully convolutional one-stage object detection[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 9626–9635.

[15] GE Z, LIU S T, WANG F, et al. YOLOX: exceeding YOLO series in 2021[DB/OL]. [2021-08-06]. https://arxiv.org/abs/2107.08430.

[16] HAN K, WANG Y H, TIAN Q, et al. GhostNet: more features from cheap operations[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 1577–1586.

[17] HOU Q B, ZHOU D Q, FENG J S. Coordinate attention for efficient mobile network design[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 13708–13717.

[18] RAO Y M, ZHAO W L, TANG Y S, et al. HorNet: efficient high-order spatial interactions with recursive gated convolutions[DB/OL]. [2022-10-11]. https://arxiv.org/abs/2207.14284.

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

[20] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc. , 2017: 6000–6010.

[21] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[DB/OL]. [2021-06-03]. http://arxiv.org/abs/2010.11929.

[22] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, et al. The Pascal visual object classes (VOC) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303−338.

[23] EVERINGHAM M, ESLAMI S M A, VAN GOOL L, et al. The Pascal visual object classes challenge: a retrospective[J]. International Journal of Computer Vision, 2015, 111(1): 98−136.

[24] ZHENG Z H, WANG P, LIU W, et al. Distance-IoU loss: faster and better learning for bounding box regression[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 12993–13000.

[25] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018: 3–19.

（编辑：董 伟）