智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于YOLOv4改进特征融合及全局感知的目标检测算法

程德强, 马尚, 寇旗旗, 张皓翔, 钱建生

引用本文:
程德强,马尚,寇旗旗,张皓翔,钱建生. 基于YOLOv4改进特征融合及全局感知的目标检测算法[J]. 智能系统学报, 2024, 19(2): 325–334.
CHENG Deqiang, MA Shang, KOU Qiqi, et al. Target detection algorithm for improving feature fusion and global perception based on YOLOv4[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(2): 325–334.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202207018

DOI: 10.11992/tis.202207018
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20231115.1446.004

基于 YOLOv4 改进特征融合及全局感知的目标检测算法

程德强[1], 马尚[1], 寇旗旗[2], 张皓翔[1], 钱建生[1]
(1. 中国矿业大学 信息与控制工程学院, 江苏 徐州 221116; 2. 中国矿业大学 计算机科学与技术学院, 江苏 徐州 221116)

摘 要: YOLOv4 算法在检测速度和精度上达到了很好的平衡, 但仍存在着定位框不准确、检测率低的问题, 尤其是在检测目标较小、尺度变化大的情况下。针对以上问题, 提出一种新的基于 YOLOv4 改进的目标检测算法。该算法采用改进的特征融合模块(path aggregation network combined with bi-directional feature pyramid network, P-Bifpn)代替 PANet(path aggregation network), 增加跨尺度连接的同时在输出端引入权重, 增强重要特征的表现力, 解决由多尺度变化而引起的精度下降。然后, 采用新的全局注意力机制(global association network, GANet), 在减少平均池化与计算量的同时增强 Sigmoid 函数输出, 加强模型对目标上下文关系的学习, 减少噪声干扰和全局信息的损失。试验采用 RSOD、NWPU VHR-10 数据集, 平均检测精度分别提升了约 5% 和 3%; 泛化试验采用 VOC2007+2012 公共数据集, 平均检测精度提升了约 0.6%。试验结果表明改进的算法能够有效提高模型的检测能力。

关键词: YOLOv4; 目标检测; 特征融合; 跨尺度; 多尺度变化; 全局注意力; 平均池化; 上下文信息
中图分类号: TP391 **文献标志码**: A **文章编号**: 1673−4785(2024)02−0325−10

中文引用格式: 程德强, 马尚, 寇旗旗, 等. 基于 YOLOv4 改进特征融合及全局感知的目标检测算法 [J]. 智能系统学报, 2024, 19(2): 325–334.
英文引用格式: CHENG Deqiang, MA Shang, KOU Qiqi, et al. Target detection algorithm for improving feature fusion and global perception based on YOLOv4[J]. CAAI transactions on intelligent systems, 2024, 19(2): 325–334.

Target detection algorithm for improving feature fusion and global perception based on YOLOv4

CHENG Deqiang[1], MA Shang[1], KOU Qiqi[2], ZHANG Haoxiang[1], QIAN Jiansheng[1]
(1. School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China; 2. School of Computer Science & Technology, China University of Mining and Technology, Xuzhou 221116, China)

Abstract: The YOLOv4 algorithm has a good balance in detection speed and accuracy, but there are still drawbacks of inaccurate positioning frame and low detection rate, especially for small detection targets and great changes in scale. Dealing with these problems, a new YOLOv4-based target detection algorithm is developed. The algorithm utilizes an enhanced feature fusion module—PANet combined with the bidirectional feature pyramid network instead of PANet to increase cross-scale connections, introduce weights at the output to improve the expressiveness of important features and solve accuracy degradation as a result of multiscale changes. Afterward, a new global association network is adopted to improve the output of the Sigmoid function while reducing the average pooling and computation, strengthen the model's learning of the contextual relationship of the target, and reduce noise interference and global information loss. The RSOD and NWPU VHR-10 datasets are employed here, with average detection accuracies being enhanced by about 5% and 3%, respectively; the generalization experiment uses the VOC2007 + 2012 public dataset, with the average detection accuracy being enhanced by about 0.6%. The experimental results reveal that the improved algorithm can effectively enhance the detection ability of the model.

Keywords: YOLOv4; target detection; feature fusion; cross-scale; multiscale variation; global attention; average pooling; contextual information

收稿日期: 2022−07−12. **网络出版日期**: 2023−11−15.
基金项目: 国家自然科学基金项目(52204177).
通信作者: 程德强. E-mail: chengdq@cumt.edu.cn.

目标检测技术是计算机视觉领域中重要的一环, 已广泛地应用于交通监控、工业生产和航空

航天等领域。当前目标检测仍然存在检测精度低等问题，尤其是在检测目标干扰因素多，尺度变化大。由于航拍图像受到光线、天气和尺度变换等因素的影响，航拍目标检测具有很大的挑战性。传统的目标检测采取 SIFT（(scale-invariant feature transform）、HOG（histogram of oriented gradient）等特征提取方法，基于图像的纹理、色彩和尺度等特征[1-3]，通过传统机器学习分类器对滑动窗口进行分类以达到检测目的。由于滑动窗口的使用时间复杂度高且产生冗余的窗口，严重影响传统检测方法的速度和精度[4]。

基于深度卷积神经网络的目标检测算法根据网络阶段可以分为 One-stage 和 Two-stage 2 种方法。Two-stage 检测算法将目标检测分为两个阶段：首先产生候选区域 (region proposals)；然后将这些区域经过神经网络进行位置微调和分类，输出最终的检测效果。这类算法基于 RCNN[5]（region with CNN feature）网络，主要包括 Fast R-CNN[6]、Faster R-CNN[7] 和 Mask R-CNN[8] 等。其检测准确率高，漏检率低，但是速度较慢，不能满足实时检测。One-stage 检测算法将回归和分类统一视作回归问题[9]，直接得到检测目标的类别概率和位置坐标值，通过端到端的网络直接输出检测结果。相对于双阶段检测算法，One-stage 有着更快的检测速度。这类算法主要包括 YOLO[10] 系列 (you only look once)、SSD[11]（single shot multibox detector）和 RetinaNet[12] 等。

2013 年，Girshick 等[5] 最先提出了卷积神经网络 RCNN。2015 年又相继提出了 Fast-RCNN[6] 和 Faster-RCNN[7] 算法。同年，Redmond 等[10] 提出了 YOLO 算法，之后 YOLO 系列算法不断发展，YOLOv2、YOLOv3[13] 算法相继问世。2020 年，Bochkovskiy 等[14] 提出了 YOLOv4 算法。该算法在 YOLOv3 算法上进行改进，融合了路径聚合网络[15](PANet)、马赛克数据增强、Mish 激活函数自对抗训练等多个模块，实现了检测速度和精度的最佳权衡。2021 年，Ge 等[16] 提出了 YOLOX 算法，该算法最大的改进是取消了在预测端（Prediction）使用多个锚框预测物体的位置和类别，使网络检测头处参数量减少 66%。由于 YOLOv4 网络过多模块的导入，导致其在提取特征时存在目标特征丢失等问题，尤其是在检测背景复杂，噪声干扰多的情况下。针对这一问题，本研究选取 RSOD[17] 和 NWPU VHR-10[18] 航拍数据集进行试验，在 VOC2007+2012 数据集进行泛化试验。航拍图像：1）图像受到天气、光线等因素影响，背景复杂，噪声干扰多；2）图像由高点拍摄，目标相对较小；3）拍摄图像过程中，容易出现目标的多尺度变换。YOLOv4 算法在该类图像上存在漏检、误检严重。针对以上问题，本研究提出了一种基于 YOLOv4 改进的特征融合及全局感知算法，主要贡献如下：1）提出新的特征融合模块 P-Bifpn（path aggregation network combined with bi-directional feature pyramid network），在 PANet（path aggregation network）基础上额外增加跨尺度连接和一条自下而上的通道，并在每个特征融合节点引入权重，关注更加重要的特征，改善由多尺度变化引起的检测精度低的问题；2）提出一个学习上下文信息的全局感知注意力网络（global association network, GANet），减少网络参数的同时获取包含全局信息的强相关性表征，捕获长远距离依赖，降低干扰因素对检测的影响，提高检测准确率。

1　相关工作

1.1　YOLOv4 算法

YOLOv4 网络主要包含以下 4 个部分：CSPDarknet53[19]（主干）、SPP 特征金字塔网络 (颈)[20]、PANet 路径聚合 (颈) 以及 YOLOv3 Head(头部)。输入一张图片，将尺寸调整为 416 像素×416 像素后通过 CSPDarknet53 网络进行特征提取，之后经过 SPP 特征金字塔模块，利用不同尺度的最大池化对上层输出的特征图进行处理。再经过 PANet 特征融合模块，将不同尺度的特征进行融合，得到 3 种不同尺度的特征图。最后，将得到的特征图划分为 $N\times N$ 个网格，并且每一个网格对应 3 个锚框，进行类别以及位置信息的预测。

1.2　多尺度特征融合

在目标检测框架中，FPN[21](feature pyramid networks) 通常用于多尺度融合，该结构采用自上而下的方式融合高分辨率和低分辨率的特征。这种融合方式有效地提高模型的性能，实现了多尺度融合。然而，该方法在特征融合过程中存在特征融合不充分等问题，导致不能对图像目标进行准确定位，导致了检测精度降低。2018 年，PANet 结构模型被提出，该模型在 FPN 基础上增加一个自下而上的路径来融合不同分辨率的特征。2020 年，谷歌团队推出 Bifpn 结构[22]。Bifpn 涉及到 2 个方面，即跨尺度连接和加权特征融合。通过重复自顶向下和自底向上的结构，融合不同尺度的特征，实现高效的特征融合，但同时也增大

了网络的模型和计算量。同年, Liu 等[23] 针对航拍图像的特征, 提出了 UAV-YOLO, 该算法在 YOLOv3 的基础上将 2 个具有相同高度和宽度的 ResNet 单元串联起来, 丰富空间信息, 扩大感受野, 但检测精度有待提高。王凤随等[24] 提出一种自适应上下文的特征融合网络 A-PANet, 通过自适应调整不同分辨率特征间的依赖性, 实现位置特征与语义特征的有效融合。赵文清等[25] 提出了双向特征融合模块, 将深层特征层进行双线性插值放大与浅层特征进行特征融合, 然后将得到的特征通过降采样的方式与深层信息进行融合。本研究主要基于 Bifpn 结构进行改进, 其初始结构如图 1 所示。

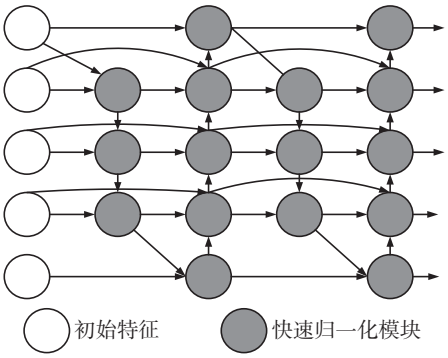


图 1　Bifpn 结构图[22]
Fig. 1　Bifpn structure[22]

1.3　注意力机制

注意力机制[26] 已经成为深度学习中研究的重要领域。注意力机制能够使网络更加关注特征的关键信息, 从而提升模型的性能。注意力机制按照不同的关注域可以分为空间域(spatial domain)、通道域(channel domain)、层域(layer domain)等。Hu 等[27] 提出了 SENet(squeeze-and-excitation network)注意力机制, 通过挤压和激励模块对特征图进行挤压操作, 获得通道级全局特征。然而, 这种注意力机制仅仅关注通道, 并未充分利用上下文信息。2018 年, Woo 等[28] 提出了 CBAM(convolutional block attention module)注意力机制, 在 SENet 的基础上在通道和空间维度上加入注意力机制。2020 年, Tian 等[29] 提出基于全卷积的单级目标检测器改进的级联注意力机制, 增强图像的语义特征。自注意力机制(self-attention)是由 Vaswani 等[30] 首次提出的, 通过利用该机制来获取输入的全局依赖性并且应用于自然语言处理中。随后, 该机制也被用于视觉图像领域, 如 2018 年, 基于自注意力机制的 Non Local 思想[31] 被引入到视频分类过程中, 在提取某处特征的同时利用其周围信息, 这个周围信息既可以是时间维度的, 也可以是空间维度的, 更好地利用时序上的信息。本研究基于自注意力机制进行改进, 其初始结构如图 2 所示。

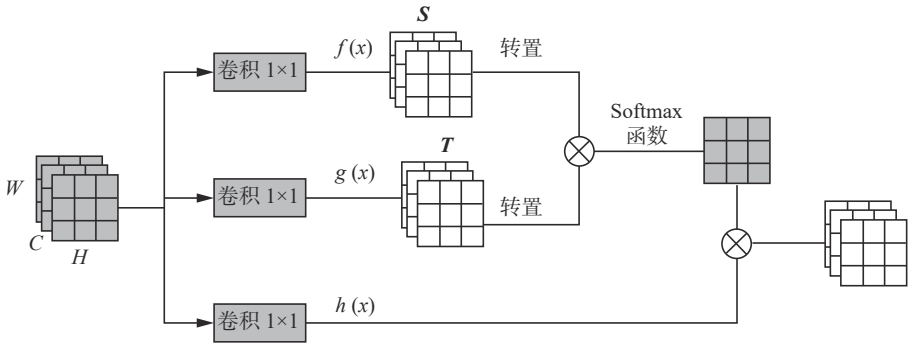


图 2　自注意力网络[30]
Fig. 2　self-attention network[30]

2　改进算法

2.1　改进的特征融合及全局感知算法

本研究在 YOLOv4 算法基础上, 提出一种新的基于加权双向特征融合及全局感知的目标检测算法, 解决目前检测中由于尺度变化大、干扰因素多而引起的检测精度低等问题。首先提出一种结合 PANet 与 Bifpn 的特征融合模块 P-Bifpn, 对提取的多尺度特征进行跨尺度连接, 同时, 为每个输入添加额外的权重, 让网络了解每个输入特征的重要性, 实现更加高效的多尺度融合, 缓解检测过程中遇到的尺度变化大的问题。其次, 本研究提出全局感知注意力网络(GANet), 利用图像的上下文信息获取相关性特征, 再利用相关性特征及原始特征进行注意力学习, 减少噪声和背景的干扰, 降低全局信息的缺失对检测精度的影响, 扩大感受野, 提高检测性能。其网络结构如图 3 所示。

由图 3 可知，本研究采用 CSPDarknet53 为网络特征提取结构，将提取的特征通过 P-Bifpn 特征融合模块，将融合后的特征通过 GANet 注意力模块，最后将得到的特征通过 YOLO Head 进行输出。

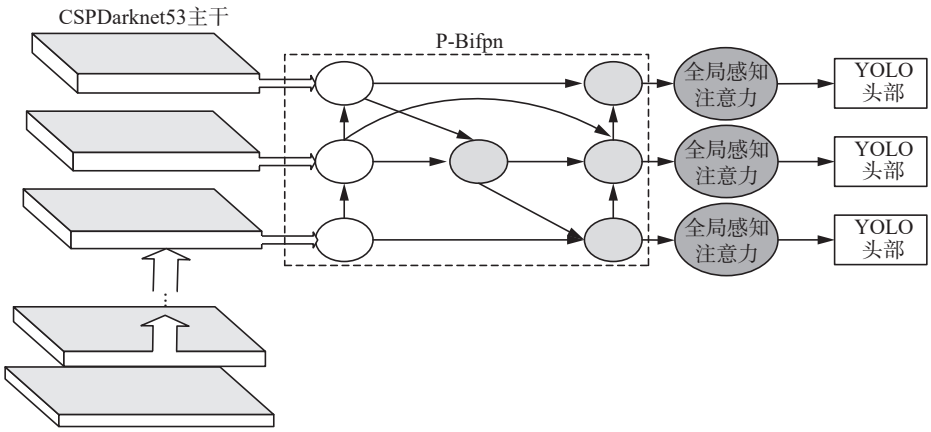


图 3　改进的 YOLOv4 网络结构

Fig. 3　Improved YOLOv4 network structure

2.2　改进的特征融合模块 (P-Bifpn)

本研究采用简化版的 Bifpn 结构，将输入特征从 5 个降为 3 个，中间的融合模块降为 1 个，输出特征为 3 个。结合 PANet 的特征模块，前 2 条路径采用 PANet 方式，在输出端采用 Bifpn 方式。其结构如图 4 所示。

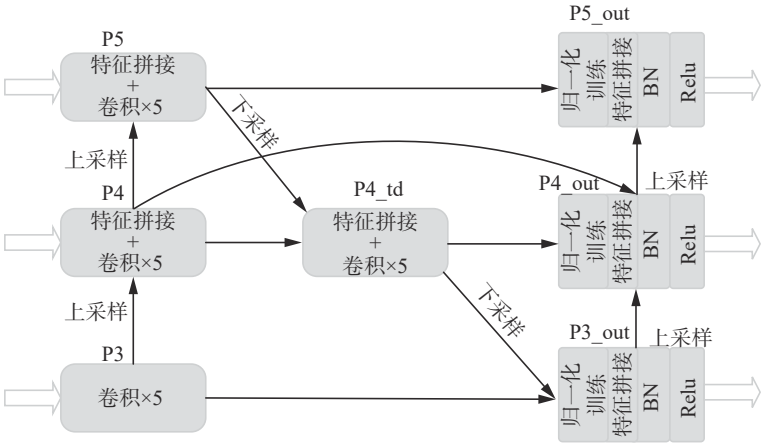


图 4　P-Bifpn 网络结构

Fig. 4　P-Bifpn network structure

图 4 中，P3、P4、P5、P4_td 为 PANet 模块，在原始基础上，增加一条 P4 指 P4_out 的残差边，融合更多尺度的特征。由于输入的特征对输出特征的贡献是不同的，增加一条由 P3_out、P4_out 以及 P5_out 组成的自下而上的通道，在有多个特征进行融合的节点引入可以训练的权重，与每个节点的输入特征相乘，从而满足不同输入特征对最终输出的不同贡献。本研究采用快速归一化公式来训练这些权重

$$o=\sum_i \frac{\omega_i}{\varepsilon+\sum_j \omega_j}\cdot I_i \tag{1}$$

式中：ω_i、ω_j代表不同特征的权重，在得到每个ω_i后引入 Relu 函数来确保$\omega_i \geqslant 0$，且ε=0.000 1，以避免数值的不确定性；I_i 代表第 i 个输入特征。

将 P3_out、P4_out 以及 P5_out 节点输入的特征代入式 (1)，得到 P3_out、P4_out 以及 P5_out。P5_out 与 P5、P4_td 进行特征融合，P4_out 与 P4、P4_td 以及 P5_out 上采样结果进行特征融合，P3_out 与 P3、P4_out 的上采样结果进行特征融合。在快速归一化训练权重后，添加批量归一化和 Relu 激活进一步提高融合效率。通过添加跨尺度连接和对输出特征自下而上的拼接，达到多尺度特征融合的目的，从而减少多尺度变化带来的精度损失。

2.3　全局感知注意力网络 (GANet)

目标检测中的物体都不是单独出现的，通常

以相同的背景或者与其他位置相关的物体一起出现。为了充分利用检测物体的背景和其他相关物体的关联性, 本研究提出一种新的全局注意力机制 GANet。全局注意力网络 GANet 类似于 Non Local 思想, 在自注意力机制的基础上进行改进。首先, 为了在减少计算量的同时不损失各个目标检测物之间的相关性, 将得到的相关性特征图不经过 Softmax 处理, 通过相关性特征映射图分为 2 个特征, 然后进行全局池化, 无需更多的矩阵乘法运算, 大大减少了计算量, 2 个相关性特征也充分表达了各个位置的相关性信息。其网络结构图如图 5 所示。

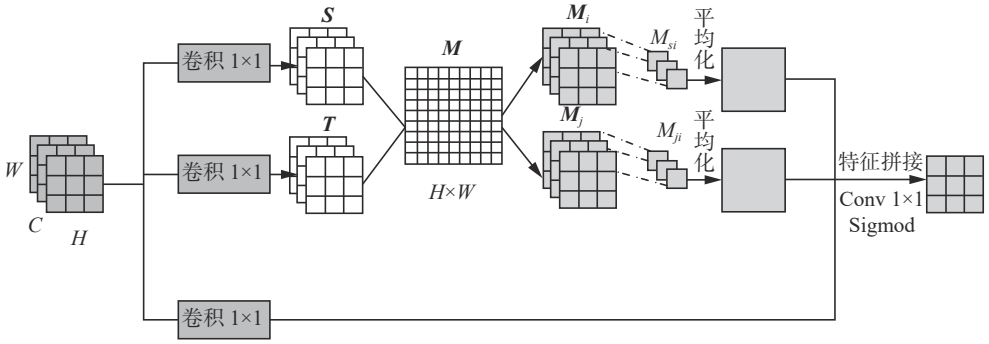


图 5　全局感知注意力网络

Fig. 5　Global aware attention network

图 5 中, $\boldsymbol{S}$、$\boldsymbol{T}$ 代表自相关性的特征图, $\boldsymbol{M}$ 代表相关特征映射图, $\boldsymbol{M}_i$ 和 $\boldsymbol{M}_j$ 代表划分出的相关性特征。

首先, 将得到的特征图通过 2 个 1×1 的卷积得到特征图 $\boldsymbol{S}$、$\boldsymbol{T}$, 其特定位置的特征值用 $\boldsymbol{s}_x$、$\boldsymbol{t}_x$ 来表示。按照顺序对特征图 $\boldsymbol{S}$ 进行扫描, 将每一个位置的特征通过自相关函数 $f_m(\cdot)$ 计算与 $\boldsymbol{T}$ 上所有特征值的相关性, 得到特征图 $\boldsymbol{M}$。由图 5 可知, $\boldsymbol{M}$ 处的每个位置都保留一个位置与另外一个位置的空间相关性, 包含了双向的相关性信息。其中 $f_m(\cdot)$ 为空间余弦相似度函数, 计算特征映射图 $\boldsymbol{M}$ 中 (x, y) 处在原始特征中 x 处与 y 处的相似度 $m(x,y)$, 其计算公式如下:

$$m(x,y) = f_m(\boldsymbol{s}_x, \boldsymbol{t}_y) = \mathrm{dot}(\boldsymbol{s}_x^{\mathrm{T}}, \boldsymbol{t}_y) \tag{2}$$

将得到的相关特征映射图 $\boldsymbol{M}$ 中的双向相关性信息分别提取出来, 得到 2 个新的相关性特征 $\boldsymbol{M}_i$ 和 $\boldsymbol{M}_j$:

$$M_{si} = M(i,:),\ \ M_{ji} = M(:,i) \tag{3}$$

最后, 再将得到的相关性特征 $\boldsymbol{M}_i$ 和 $\boldsymbol{M}_j$ 与初始特征进行拼接, 拼接后的特征经过 Conv 1×1 与 Sigmoid 函数输出注意力权重值, 来更好地学习全局感知注意力。

2.4　损失函数

YOLO 网络将检测问题化为回归问题, 生成每个类的边界坐标和概率。如果被检测物体中心落在网格内, 网格将根据人工标记的区域进行目标检测, 通过边界盒位置损失 L_{CIoU}(bounding box location loss)、置信度损失 $L_{\mathrm{confidence}}$(confidence loss)和分类损失 L_{class}(classification loss)对 YOLOv4 损失函数进行训练

$$L_{\mathrm{total}} = L_{\mathrm{CIoU}} + L_{\mathrm{confidence}} + L_{\mathrm{class}} \tag{4}$$

$$C_{\mathrm{IoU}} = I_{\mathrm{IoU}} - \frac{\rho^2(b,b^{\mathrm{gt}})}{c^2} - \alpha v \tag{5}$$

式中: I_{IoU} 是定义目标检测精度的标准, 表示预测边界框与地面真值边界框的交集比; $\rho^2(b,b^{\mathrm{gt}})$ 表示预测帧中心点和真实帧之间的欧几里得距离; c 表示可以同时包含预测帧和真实帧的最小所需区域的对角线距离; α 代表权重参数; v 用来评定纵横比的一致性

$$\alpha = \frac{v}{1 - I_{\mathrm{IoU}} + v} \tag{6}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{\mathrm{gt}}}{h^{\mathrm{gt}}} - \arctan\frac{\mathrm{w}}{h}\right) \tag{7}$$

$$L_{C_{\mathrm{IoU}}} = 1 - I_{\mathrm{IoU}} + \frac{\rho^2(b,b^{\mathrm{gt}})}{c^2} + \alpha v \tag{8}$$

式(7)中, w^{gt} 和 h^{gt} 为真实框的宽、高; w 和 h 表示预测框的宽、高。

3　试验

3.1　数据集

为了验证本研究算法的有效性, 采用航拍图像数据集 RSOD 和 NWPU VHR-10 作为试验对象。RSOD 和 NWPU VHR-10 是用于航拍检测和空间物体检测的数据集, 其图像尺度变化大, 一些检测目标较小, 且干扰因素较多, 满足本研究的试验要求。泛化试验采用 VOC2007+2012 数据集。

RSOD 数据集包含 4 类, 共 936 张图片, 大约在 1 000 像素×900 像素~1 200 像素×1 000 像素, 检测目标尺寸变化较大且分布不均。其中, 飞机类目标较小, 分布密集, 检测目标多; 油桶类尺寸适

中，分布较密；立交桥和操场为目标较大，且分布较少。将该数据集按训练集、验证集、测试集为6∶2∶2划分。

NWPU VHR-10 数据集包含 10 个类别，共 650 张图片，大约在 800 像素×500 像素~1 000 像素×500 像素。该数据集种类较多，目标尺寸变化大，分布较密集。将该数据集按训练集、验证集、测试集为 8∶1∶1 划分。

VOC2007+2012 数据集包含 20 个类别，共 11 530 张图片，将该数据集按训练集、验证集、测试集为 8∶1∶1 划分。

3.2　试验设置

本试验采用 CPU 型号：Intel(R) Core(TM) i9-10980XE @ 3.0 GHz，内存：64 GB 显卡：2 张 GTX3090 的 GPU,显存：24 GB。在 Pytorch1.10.0 框架下运行。本试验在各数据集上 epoch 均设置为 100，batch size 为 4，RSOD 和 VOC 数据集上初始学习率为 0.000 1，NWPU VHR-10 数据集上初始学习率设置为 0.001。采用余弦退火衰减策略。

3.3　试验结果与分析

表 1 给出各个算法在 RSOD 数据集与本研究算法的对比试验结果。

表 1　不同算法在 RSOD 数据集上的检测精度对比

Table 1　Comparison of detection accuracy of different algorithms on RSOD dataset　%

类别	Faster-RCNN[7]	SSD[11]	Retina Net[12]	YOLOv3[13]	Efficientdet-D2[22]	UAV-YOLO[23]	YOLOX[16]	YOLOv4[14]	本研究算法
飞机	42.23	48.41	70.92	71.60	60.30	74.68	90.11	73.09	78.05
油桶	89.31	87.51	79.45	87.81	90.65	74.20	88.02	95.32	97.88
立交桥	97.42	70.68	35.29	65.56	89.31	76.32	50.11	65.90	74.04
操场	90.95	89.75	97.92	85.25	90.65	85.96	74.89	89.03	95.18
mAP	79.87	74.09	70.73	77.55	82.72	77.79	75.78	80.84	86.29

由表 1 可见，本研究所提出的算法在各类上的平均准确率（Average Pression，AP）、mAP 几乎全部高于其他算法。单阶段目标检测算法 Faster-RCNN 虽然在 AP、mAP 上有不错的表现，但是对于小目标类别检测准确率较低，不能满足实时检测的目标。SSD、RetinaNet 以及 YOLOv3 在检测准确率不断提高，但明显弱于 YOLOv4 算法。Efficientdet-D2 算法在飞机类上表现较差，但在其他 3 类中表现较好，可见该算法在小目标上检测性能差，在中、大目标上表现较好。UAV-YOLO 算法在所有类别上表现的都较为稳定，但是总体效果不如本研究提出的算法。YOLOX 算法在小目标检测上表现的尤为突出，达到 90%，远超其他算法，但在中大目标检测上表现较差，尺度变化对该算法影响较大，mAP 比本研究算法相差约 10%。本研究提出的算法相对于初始的 YOLOv4 算法飞机、立交桥以及操场这 3 个类别有较大的提升。其中飞机为较小的目标，且分布较密，本研究算法在该类上 AP 提高了约 5 个百分点，可见本研究算法对于小目标检测也有一定的提升能力。

由表 2 可见，本研究提出的算法 mAP 达到了近 94%，为所有算法中最高。本研究算法在中大目标上表现的效果较好且相对稳定，多个特征明显的类别 AP 均达到近 100%。在 habor、brigde、vehicle 3 个特征较弱的类别中，YOLOX 表现效果较差，bridge 类的 AP 为 39%，可见该算法对于特征较弱的中大型目标检测效果较差。YOLOv3、UAV-YOLO、YOLOv4 算法明显优于其他算法，且 3 类表现的都较为稳定。本研究算法相对于 YOLOv4 算法几乎所有类的 AP 均有所提升，mAP 提升了 3.5 个百分点，达到理想效果，证明本研究算法能够有效提高在航拍数据集上的检测能力。

表 2　不同算法在 NWPU VHR-10 数据上的检测精度对比

Table 2　Comparison of detection accuracy of different algorithms on NWPU VHR-10 dataset　%

类别	Faster-RCNN[7]	SSD[11]	Retina Net[12]	YOLOv3[13]	Efficientdet-D2[22]	UAV-YOLO[23]	YOLOX[16]	YOLOv4[14]	本研究算法
airplane	97.83	98.35	99.56	98.59	100.00	99.01	99.19	99.98	99.99
ship	78.66	71.02	78.16	84.75	70.37	84.60	78.58	83.95	85.49
storage tank	90.68	80.35	82.29	99.98	78.19	99.98	99.97	100.00	100.00
baseball diamond	89.99	88.40	99.55	97.41	98.71	98.12	98.94	97.90	99.50
tennis court	80.85	89.27	83.37	99.01	89.60	98.99	93.28	99.09	99.10

续表 2

类别	Faster-RCNN[7]	SSD[11]	Retina Net[12]	YOLOv3[13]	Efficientdet-D2[22]	UAV-YOLO[23]	YOLOX[16]	YOLOv4[14]	本研究算法
basketball court	58.80	70.91	65.18	90.32	94.13	90.30	73.72	99.31	99.19
groundtrack field	95.47	99.45	95.38	95.37	100	96.25	79.15	95.45	99.62
habor	80.68	85.94	65.66	82.99	78.95	80.65	64.50	89.24	87.21
bridge	63.33	63.92	40.35	75.33	53.83	72.06	38.99	71.13	74.14
vehicle	73.09	54.61	71.91	82.04	75.88	81.96	81.71	92.63	94.49
mAP	80.94	80.22	78.13	90.57	83.97	90.19	80.81	91.21	93.87

由表 3 可知, 本研究在公共数据集 VOC2007+12 上进行泛化试验。该数据集包含 20 个类别、1 万多张图片以及 27 000 多标注物体, 能够有效验证本研究算法的有效性。表 3 中, Faster-RCNN 在 mAP 与检测速度上均为最低, 为 73.28% 与 5.1 帧/s, SSD 检测速度最快, 但在精度上表现较差。UAV-YOLO 相对于 YOLOv3 检测精度和速度都有所上升, 但是整体提升不大。YOLOX 算法平均检测精度达到 86.84%, 且检测速度也较高。该算法相对于 YOLOv4, 在输出端使用无锚机制, 只预测一个后选框, 同时引入 Mixup、多尺度增强等, 在大大减少计算量的同时大大提高检测精度, 相对于本研究算法, 该算法并没有充分考虑到全局信息与多尺度变化, 在 YOLOv4 引入自注意力机制后, mAP 有了显著的提升, 且在航拍数据集中, 由于多尺度变化、背景噪声等对 YOLOX 影响较大。本研究算法虽然在检测速度上有一定降低, 但检测精度在所有对比试验中为最高, 证明了本研究算法的有效性。

表 3　在 VOC2007+12 上与主流算法对比

Table 3　Comparisons with state-of-art methods on VOC2007+12

算法	mAP/%	检测速度/(帧·s^{-1})	算法	mAP/%	检测速度/(帧·s^{-1})
Faster-RCNN[7]	73.28	5.1	UAV-YOLO[23]	81.83	12.8
SSD[11]	79.80	19.8	YOLOX[16]	86.84	13.0
RetinaNet[12]	81.56	9.6	YOLOv4[14]	87.35	10.0
YOLOv3[13]	81.63	13.4	YOLOv4+self-attention	87.89	9.0
Efficientdet-D2[22]	82.95	6.3	本研究算法	88.01	8.4

3.4　消融试验

为了验证本研究提出算法的有效性, 针对提出的改进的特征融合模块(P-Bifpn)和 GANet 进行消融试验。该试验在 RSOD 数据集上进行。

P-Bifpn 模块消融试验效果如表 4 和表 5 所示。

表 4　P-Bifpn 模块消融试验

Table 4　Albation studies on P-Bifpn

算法	mAP/%	检测速度/(帧·s^{-1})
YOLOv4[14]	80.84	10.0
YOLOv4+Bifpn	79.81	10.6
YOLOv4+Bifpn^{+}	81.67	9.6
YOLOv4+P-Bifpn	83.11	8.9

表 4 中, YOLOv4+Bifpn 代表用 Bifpn 网络完全代替 PANet 网络, 该方法减少了模型的计算量, 提高了检测速度, 但是检测精度下降。YOLOv4+Bifpn^{+}表示先采用 PANet 网络结构, 然后在输出端采用快速归一化加权融合的特征融合方式, 虽然该模块检测速度较高, 但是 mAP 较 P-Bifpn 相差 1.5%。本研究提出的 P-Bifpn 模块在输出端采用特征拼接的方式, 在增加少量计算量的同时提高检测精度。全局感知注意力机制 GANet 消融试验如表 5 所示。

表 5　GANet 模块消融试验

Table 5　Albation studies on GANet

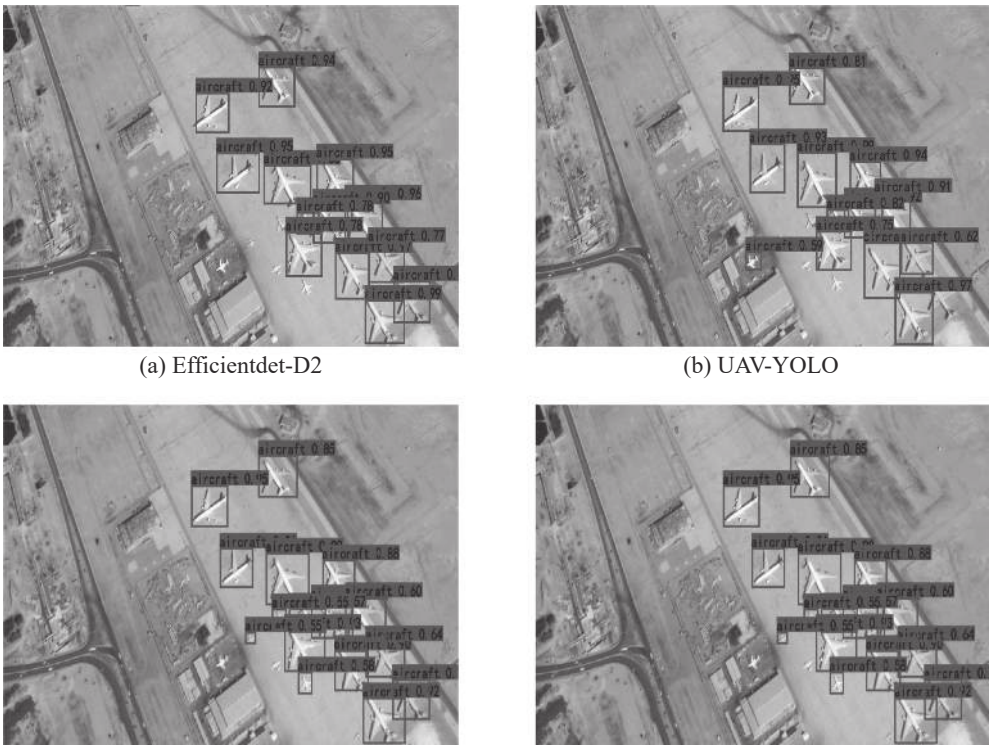
算法	mAP/%	检测速度/(帧·s^{-1})
YOLOv4[14]	80.84	10.0
YOLOv4+SENet	81.70	9.5
YOLOv4+GANet-	82.93	9.4
YOLOv4+GANet	84.48	9.0

本研究在 YOLOv4 基础上首先引入自注意力机制网络, 其检测精度有了显著提升。为了减少模型的参数量, 对 2 个相关性特征 $\boldsymbol{M}_i$ 和 $\boldsymbol{M}_j$ 进行通道级平均池化, 再与原特征融合学习注意力权重。由于通道级平均池化会造成严重的特征语义丢失, 故采用 1×1 卷积代替, 得到 GANet-, 检测精度提升了约 1%。通过试验发现利用 Sigmoid 函数输出的注意力权重较小, 因此将 Sigmoid 函数

输出的注意力权重加到原始特征上，有效地缓解由于权重太小造成对特征值的整体削弱，得到本研究提出的 GANet，相对于 YOLOv4 提升了模型的检测精度约 4%。

3.5　试验结论

上述在 RSOD 和 NWPU VHR-10 数据集上进行的对比试验和消融试验结果显示，本研究提出的算法在针对航拍类数据集有一定优势：首先，本研究提出的改进的特征融合模块，在 RSOD 数据集上大大提高了分布密集的小目标的检测精度，并且具有良好的泛化性；本研究提出的 GANet 网络结构对于模型精度也有较大改善。最终，本研究提出的算法较其他算法有一定的优越性。其可视化结果如图 6 所示。



(a) Efficientdet-D2　(b) UAV-YOLO

(c) YOLOv4　(d) 本文算法

图 6　不同算法的可视化结果

Fig. 6　Visualization results of different algorithms

4　结束语

本研究针对 YOLOv4 在检测过程中由目标尺度变化大、噪声干扰多和检测背景复杂等引起的检测精度低的问题，提出了一种基于改进的特征融合及全局感知的目标检测算法。该算法利用 P-Bifpn 模块进行特征融合，采用跨尺度连接，融合更多尺度的信息，同时，为每个输出特征引入额外的可训练的权重，对重要的特征进行更多的关注，实现高效的多尺度融合。然后，利用改进的全局感知注意力机制，减少计算量的同时充分学习检测目标的上下文信息，减少噪声及背景对检测的干扰，提高检测性能。经试验表明，本研究所提出的算法有效提高了模型的检测能力，且具有良好的泛化性。

未来将会进一研究对于该模型的优化，尝试代替 CSPDarknet 主干网络，减小模型的大小，提高模型的检测速度，并更好地应用于各领域实时场景检测。

参考文献：

[1]　程德强, 李腾腾, 郭昕, 等. 改进的 SIFT 邻域投票图像匹配算法 [J]. 计算机工程与设计, 2020, 41(1): 162–168.
CHENG Deqiang, LI Tengteng, GUO Xin, et al. Improved SIFT neighborhood voting image matching algorithm[J]. Computer engineering and design, 2020, 41(1): 162–168.

[2]　CHENG Deqiang, TANG Shixuan, FENG Chenchen, et al. Extended HOG-CLBC for pedstrain detection[J]. Optoelectronic engineer, 2018, 45(8): 180111.

[3]　张桂梅, 张松, 储珺. 一种新的基于局部轮廓特征的目标检测方法 [J]. 自动化学报, 2014, 40(10): 2346–2355.
ZHANG Guimei, ZHANG Song, CHU Jun. A new object detection algorithm using local contour features[J].

Acta automatica sinica, 2014, 40(10): 2346–2355.

[4] 王彦情, 马雷, 田原. 光学遥感图像舰船目标检测与识别综述 [J]. 自动化学报, 2011, 37(9): 1029–1039.
WANG Yanqing, MA Lei, TIAN Yuan. State-of-the-art of ship detection and recognition in optical remotely sensed imagery[J]. Acta automatica sinica, 2011, 37(9): 1029–1039.

[5] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. New York: ACM, 2014: 580−587.

[6] CHE Xiangjiu, LIU Hualuo, SHAO Qingbin. Fabric defect recognition algorithm based on improved Fast RCNN[J]. Journal of Jilin University (Engineering and Technology Edition), 2019, 49(6): 2038–2044.

[7] 黄继鹏, 史颖欢, 高阳. 面向小目标的多尺度 Faster-RCNN 检测算法 [J]. 计算机研究与发展, 2019, 56(2): 319–327.
HUANG Jipeng, SHI Yinghuan, GAO Yang. Multi-scale faster-RCNN algorithm for small object detection[J]. Journal of computer research and development, 2019, 56(2): 319–327.

[8] SONG Ling, XIA Zhimin. Research on improved mask R-CNN network model for human keypoint detection[J]. Computer engineering and applications, 2021, 57(1): 150–160.

[9] 刘学平, 李玓乾, 刘励, 等. 自适应边缘优化的改进 YOLOV3 目标识别算法 [J]. 微电子学与计算机, 2019, 36(7): 59–64.
LIU Xueping, LI Yuqian, LIU Li, et al. Improved YOLOV3 target recognition algorithm for adaptive edge optimization[J]. Microelectronics & computer, 2019, 36(7): 59–64.

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

[11] LIU Wei, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector[C]//European Conference on Computer Vision. Cham: Springer, 2016: 21−37.

[12] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 42(2): 318–327.

[13] REDMON J, FARHADI A. YOLOv3: an incremental improvement[EB/OL]. (2018−04−08)[2022−01−01]. https://arxiv.org/abs/1804.02767.

[14] BOCHKOVSKIY A, WANG Chienyao, LIAO Hongyuan. YOLOv4: Optimal Speed and Accuracy of Object Detection[EB/OL]. (2020−04−23)[2022−01−01]. https://arxiv.org/abs/2004.10934.

[15] LIU Shu, QI Lu, QIN Haifang, et al. Path aggregation network for instance segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 8759−8768.

[16] GE Zheng, LIU Songtao, WANG Feng, et al. YOLOX: exceeding YOLO series in 2021[EB/OL]. (2021−07−18)[2022−01−01]. https://arxiv.org/abs/2107.08430.pdf.

[17] ZHANG Xiangyu, ZHOU Xinyu, LIN Mengxiao, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 6848−6856.

[18] CHENG Gong, HAN Junwei, ZHOU Peicheng, et al. Multi-class geospatial object detection and geographic image classification based on collection of part detectors[J]. ISPRS journal of photogrammetry and remote sensing, 2014, 98: 119–132.

[19] AGGARWAL V, WANG Wenlin, ERIKSSON B, et al. Wide compression: tensor ring nets[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 9329−9338.

[20] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904–1916.

[21] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 936−944.

[22] TAN Mingxing, PANG Ruoming, LE Q V. EfficientDet: scalable and efficient object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 10778−10787.

[23] LIU Mingjie, WANG Xianhao, ZHOU Anjian, et al. UAV-YOLO: small object detection on unmanned aerial vehicle perspective[J]. Sensors, 2020, 20(8): 2238.

[24] 王凤随, 陈金刚, 王启胜, 等. 自适应上下文特征的多尺度目标检测算法 [J]. 智能系统学报, 2022, 17(2): 276–285.
WANG Fengsui, CHEN Jingang, WANG Qisheng, et al. Multi-scale target detection algorithm based on adaptive

context features[J]. CAAI transactions on intelligent systems, 2022, 17(2): 276–285.

[25] 赵文清, 杨盼盼. 双向特征融合与注意力机制结合的目标检测 [J]. 智能系统学报, 2021, 16(6): 1098–1105.
ZHAO Wenqing, YANG Panpan. Target detection based on bidirectional feature fusion and an attention mechanism[J]. CAAI transactions on intelligent systems, 2021, 16(6): 1098–1105.

[26] WANG Hao, WANG Qilong, GAO Mingqi, et al. Multi-scale location-aware kernel representation for object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 1248−1257.

[27] HU Jie, SHEN Li, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 42(8): 2011–2023.

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

[29] TIAN Zhuoyu, MA Miao, YANG Kaifang. Object detection model for examination classroom based on cascade attention and point supervision mechanism[J]. Journal of software, 2022, 33(7): 2633–2645.

[30] 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. New York: ACM, 2017: 6000−6010.

[31] WANG Xiaolong, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7794−7803.

作者简介：

程德强，教授，博士生导师，博士，主要研究方向为计算机视觉与模式识别、图像智能检测。主持国家自然科学基金项目 3 项，江苏省重大成果转化项目等省部级各类科技项目 10 余项。以第一作者（通信作者）发表学术论文 70 余篇。E-mail：chengdq@cumt.edu.cn。

马尚，硕士研究生，主要研究方向为图像处理与目标检测。E-mail：710584238@qq.com。

寇旗旗，讲师，主要研究方向为视频、图像处理与模式识别。E-mail：137156449@qq.com。