基于 Dark-YOLO 的低照度目标检测方法

江泽涛[1)], 肖芸[1)*], 张少钦[2)], 朱玲红[3)], 何玉婷[1)], 翟丰硕[1)]

[1)] (桂林电子科技大学广西图像图形与智能处理重点实验室　桂林　541004)
[2)] (南昌航空大学土木建筑学院　南昌　330063)
[3)] (南昌航空大学信息工程学院　南昌　330063)
(zetaojiang@guet.edu.cn)

摘　要: 在复杂的低照度环境中获取的图像存在亮度低、噪声多和细节信息丢失等问题, 直接使用通用的目标检测方法无法达到较为理想的效果. 为此, 提出低照度目标检测方法——Dark-YOLO. 首先, 使用 CSPDarkNet-53 骨干网络提取低照度图像特征, 并提出路径聚合增强模块以进一步增强特征表征能力; 然后, 设计金字塔平衡注意力模块捕获多尺度特征并加以有效利用, 生成包含不同尺度且更具判别力的特征; 最后, 使用预测交并比(intersection over union, IoU)改进检测头, IoU 预测分支为每个预测框预测 IoU 值, 使得目标定位更加准确. 在 ExDark 数据集上的实验结果表明, 相较于 YOLOv4, 均值平均精度(mAP)提升了 4.10%, Dark-YOLO 方法能够有效地提高在低照度场景下目标检测的性能.

关键词: 目标检测; 低照度图像; 注意力机制; 多尺度特征; 预测交并比
中图法分类号: TP391.41　　**DOI:** 10.3724/SP.J.1089.2023.19354

Low-Illumination Object Detection Method Based on Dark-YOLO

Jiang Zetao[1)], Xiao Yun[1)*], Zhang Shaoqin[2)], Zhu Linghong[3)], He Yuting[1)], and Zhai Fengshuo[1)]

[1)] (*Guangxi Key Laboratory of Image and Graphics Intelligent Processing, Guilin University of Electronic Technology, Guilin*　541004)
[2)] (*School of Civil and Architecture, Nanchang Hangkong University, Nanchang*　330063)
[3)] (*School of Information Engineering, Nanchang Hangkong University, Nanchang*　330063)

Abstract: The images acquired in a complex low-illumination environment have problems such as low brightness, high noise, and loss of detailed information. The general object detection method cannot be used directly to achieve relatively ideal results. In this situation, a low-illumination object detection method-Dark-YOLO is proposed. Firstly, the CSPDarkNet-53 backbone network is used to extract the features of low-illumination image, and the path aggregation enhanced module is proposed to further enhance the ability of feature representation. Then, the pyramid balanced attention module is designed to capture multi-scale features and make use of them effectively to generate more discriminant features with different scales. Finally, the prediction intersection over union (IoU) is used to improve the performance of detection head. The IoU prediction branch predicts the IoU value for each prediction box, which makes the object positioning more accurate. The experimental results on the ExDark dataset show that compared with YOLOv4, the mean average precision (mAP) is improved by 4.10%. Dark-YOLO method can effectively improve the performance of object detection in low-illumination scenes.

收稿日期: 2021-09-10; 修回日期: 2021-11-14. **基金项目**: 国家自然科学基金(62172118, 61876049); 广西自然学科基金重点项目(2021GXNSFDA196002); 广西图像图形智能处理重点实验项目(GIIP2008); 广西研究生教育创新计划资助项目(YCBZ2021070, YCBZ2018052, 2020YCXS050). **江泽涛**(1961—), 男, 博士, 教授, 博士生导师, 主要研究方向为图像处理、计算机视觉; **肖芸**(1997—), 女, 硕士研究生, 论文通信作者, 主要研究方向为计算机视觉; **张少钦**(1962—), 女, 博士, 教授, 硕士生导师, 主要研究方向为优化计算; **朱玲红**(1983—), 女, 硕士, 讲师, 主要研究方向为图像处理; **何玉婷**(1995—), 女, 博士研究生, 主要研究方向为图像理解; **翟丰硕**(1997—), 男, 硕士研究生, 主要研究方向为计算机视觉.

Key words: object detection; low-illumination image; attention mechanism; multi-scale feature; prediction intersection over union

目标检测是计算机视觉领域中最具挑战的任务之一, 随着深度学习的迅速发展, 基于深度学习的目标检测方法在目标检测中占据主导地位. 通用的目标检测方法大多是在简单、干净的图像上进行训练的, 并在通用的目标检测数据集上取得了较好的检测效果. 然而, 在光照条件不足的情况下, 捕获的低照度图像往往存在亮度低、噪声多和细节信息丢失等问题, 将通用的目标检测方法直接应用于低照度图像, 会导致检测精度较低、边界框定位不准确. 低照度目标检测对夜间监控、夜间辅助驾驶、无人机夜间侦察与攻击、机器人夜间工作等场合具有大量潜在的应用前景, 因此, 设计一种高效、准确的低照度目标检测方法具有重要意义.

近年来, 随着深度学习的发展, 研究者提出了使用卷积神经网络(convolutional neural network, CNN)来提取特征. 基于深度学习的目标检测方法不需要使用复杂的手工设计特征就可以有效地提取低级特征和高级特征, 因此, 该方法比基于传统图像处理的目标检测方法在检测精度和速度方面均有极大的提升. 当前, 基于深度学习的目标检测方法主要分为 2 类: 双阶段目标检测方法和单阶段目标检测方法. 其中, 双阶段目标检测方法通过区域建议网络产生目标候选框, 在候选框的基础上进一步计算目标的分类得分并回归边界框, 检测速度较慢但检测精度较高. 其中代表性方法有 Fast R-CNN[1], Faster R-CNN[2], Mask R-CNN[3]等. 单阶段目标检测方法是直接在输入图像上密集采样不同尺度和比例的先验框, 通过匹配先验框直接定位目标, 存在检测速度较快但检测精度较低的问题. 其代表性方法有 YOLO(you only look once)系列[4-7]和 SSD(single shot multibox detector)[8]等.

基于深度学习的目标检测方法被广泛关注并走向实用, 但针对低照度场景的目标检测方法的研究较少. 对于 RGB 图像, Loh 等[9]提出了针对低照度目标检测的专用数据集 ExDark, 其对手工制作和学习特征的可视化效果进行分析, 希望能够促进低照度目标检测相关研究. 对于 RAW 图像, Sasagawa 等[10]设计 YOLO-in-the-Dark 网络模型运用粘合层和生成模型合并不同域中的预训练模型, 即低照度图像增强模型 Learning-to-See-in-the-Dark[11]和目标检测模型 YOLO[4], 生成模型将潜在特征提供给粘合层, 以在不需要额外数据集的情况下对整个模型进行训练.

针对低照度场景下由于光照不均、噪声干扰、细节信息丢失等导致检测精度较低和边界框定位不准确的问题, 本文提出 Dark-YOLO 方法用于低照度目标检测. 首先, 该方法使用特征表征能力更强的 CSPDarkNet-53[12]骨干网络进行特征提取并提出路径聚合增强模块(path aggregation enhanced module, PAEM)增强特征表征能力; 然后, 设计了金字塔平衡注意力模块(pyramid balanced attention module, PBAM)提升模型性能; 最后, 增加 IoU 预测分支改进检测头, 使得低照度场景下目标定位更准确. 本文方法实现了端到端的低照度目标检测, 在低照度目标检测数据集 ExDark[9]上进行实验验证, 具有较高的检测精度和较好的检测效果.

1 相关工作

1.1 单阶段目标检测方法

当前, 主流的单阶段目标检测方法基于回归思想, 直接对图像预测类别概率值和边界框偏移量得到分类和回归结果, 检测速度较快但检测精度较低. YOLO[4]中舍弃双阶段方法的区域建议网络, 在一个端到端的网络中实现特征提取、目标分类和回归. SSD[8]中以 VGG16[13]作为特征提取网络, 增加额外的卷积层用于检测不同大小的目标. YOLOv2[5]在速度和精度方面有更好的平衡, 提出 DarkNet-19 作为模型的骨干网络, 用于提取特征, 并减少计算量. YOLOv3[6]中将具有残差块的 DarkNet-53 作为骨干网络, 利用特征金字塔网络(feature pyramid network, FPN)[14]融合不同尺度的特征图用于检测目标. CenterNet[15]在基于关键点的单阶段目标检测方法的基础上, 将目标检测任务视为检测 3 个关键点. YOLOv4[7]中运用 CSPDarkNet-53[12]作为骨干网络, 并结合空间金字塔池化(spatial pyramid pooling, SPP)[16]、路径聚合网络(path aggregation network, PANet)[17]等先进技术, 使得网络模型达到较优的检测结果. EfficientDet[18]采用 EfficientNet[19]作为骨干网络, 提出一种加权双向特征金字塔网络(weighted bi-directional feature pyramid network, BiFPN), 使用复合缩放方法实现在资源限制下达到较好的检测精度和更小的计算量. YOLOF[20]使用膨胀编码器和统一匹配, 使得模型性能得到较大的提升.

1.2　YOLOv4

YOLOv4[7]是平衡精度和速度的单阶段目标检测方法，它可以具体拆分为 3 个部分，分别为用于特征提取的骨干网络(Backbone)、用于特征融合的颈部(Neck)和用于分类预测和边界框回归的检测头(Head). 其中，Backbone 运用 CSPNet[12]的设计思想构建全新的 CSPDarkNet-53[12]，以提高提取特征的表征能力，同时减小模型的运算量. Neck 在使用 PANet[17]进行特征融合的同时，引入 SPP[16]改善感受野大小. PANet 在 FPN[14]的基础上，增加自底向上路径来传递低级特征，以保证输出的 3 个尺度的特征图既包含底层特征又包含语义特征. SPP 运用不同尺寸的池化核对特征图进行最大池化操作，再与原特征图进行拼接输出. Neck 输出 3 个不同尺度的特征图以适应对不同大小的目标检测，Head 对 Neck 的输出分别进行卷积操作，得到最终的检测结果. YOLOv4 结构图如图 1 所示，YOLOv4 选择 CSPDarkNet-53 作为 Backbone，SPP 和 PANet 作为 Neck，以及 YOLOv3 Head 构成目标检测网络模型.

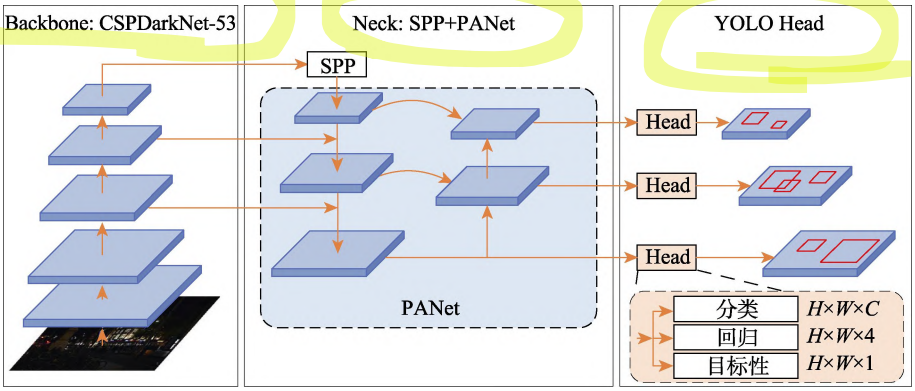


图 1　YOLOv4 结构图

1.3　注意力机制

计算机视觉领域中的注意力机制[21]是模拟人脑注意力机制的模型，其基本思想是让网络学会同人脑一样关注重点信息、忽略无关信息. 注意力机制通过为输入特征图按对输出结果的影响的重要程度分配不同的权重，以标注出有用信息并忽略无关信息. 在计算机视觉任务中运用注意力机制通过计算注意力图实现，将输入特征图送入构建的注意力模块中，通过学习注意力权重得到注意力图，最后将输入特征图与注意力图进行相乘获得输出. 注意力机制有助于计算机视觉的各项任务，如图像分类、图像分割、目标检测和图像增强等. Hu 等[22]提出 SENet 图像分类网络模型，通过对特征图的空间维度进行压缩，经过多层感知机学习注意力权重，标注出对结果贡献较大的通道进而提升模型分类的准确性. CBAM[23]通过有效地结合通道和空间注意力对输入特征进行标注，以进行自适应特征细化. 坐标注意力(coordinate attention)[24]将通道注意力分解为 2 个一维特征编码过程，分别沿 2 个空间方向聚集特征. Wang 等[25]提出轻量高效的平衡注意力机制(balanced attention mechanism, BAM)，其被应用于单幅图像超分辨率重建[26]任务.

2　Dark-YOLO

2.1　Dark-YOLO 概述

Dark-YOLO 方法中设计了 PAEM 和 PBAM，并引入预测 IoU 改进 Head. Dark-YOLO 结构如图 2 所示. Dark-YOLO 方法运用 CSPDarkNet-53 提取低照度图像特征，CSPNet[12]能够获取更丰富、强大的特征表征并减少计算量. 首先，Backbone 输出的特征图送入 PAEM，在自顶向下路径中使用空洞空间金字塔池化模块(atrous spatial pyramid pooling module, ASPPM)丰富特征，再与自底向上路径的输出特征运用加权特征融合(weighted feature fusion, WFF)进一步增强特征表征能力；其次，经过 PBAM 捕获多尺度特征，并有效地结合空间和通道注意力，让网络生成包含不同尺度且更具判别力的特征，减少噪声干扰的影响，以提升模型的检测性能；最后，引入预测 IoU 改进 Head，由 IoU 预测分支为每个预测框预测 IoU 值，将预测得到的 IoU 得分乘以目标性得分和分类得分作为预测框的置信度，使得低照度场景下目标定位更加准确.

2.2　PAEM

为进一步增强特征的表征能力，在 PANet[17]的基础上提出如图 3 所示 PAEM. PAEM 使用 ASPPM

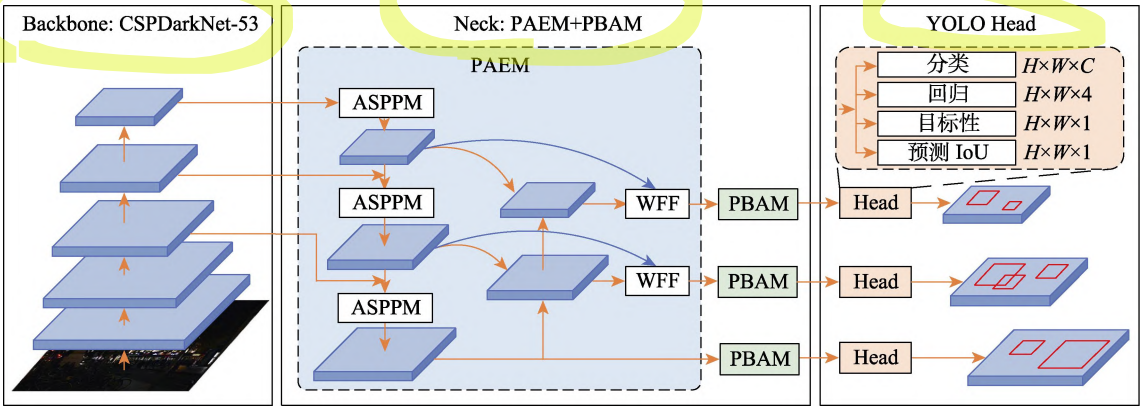


图 2 Dark-YOLO 结构图

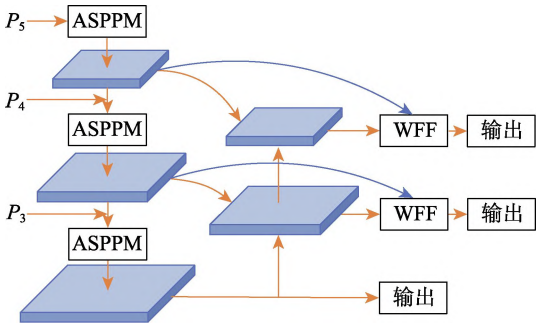


图 3 PAEM 结构图

以不同空洞率的空洞卷积提取多尺度特征；并运用 WFF 对使用 ASPPM 的特征与自底向上路径的输出特征进行有效的融合，以增强特征表征能力.

PAEM 具体运算过程：首先，将经过 Backbone 提取到的不同尺度特征图 P_3, P_4, P_5 送入 PAEM，并在自顶向下路径中运用 ASPPM 增大感受野，以提取多尺度特征；然后，通过自底向上路径传递低级特征，并将 ASPPM 输出的特征与自底向上路径的输出特征运用 WFF 得到输出特征图，以达到增强特征表征能力的目的.

2.2.1 ASPPM

在图像分割中，空洞空间金字塔池化[27]通过增大感受野来捕获上下文信息，以利于对图像每个像素分类. 受此启发，本文设计 ASPPM 应用于目标检测任务. 空洞卷积能够增大感受野而不引入更多的参数，ASPPM 用不同空洞率的空洞卷积并行地对特征图进行卷积，从而在多尺度上捕获上下文信息. ASPPM 结构图如图 4 所示.

ASPPM 采用 4 个并行的分支对输入特征图进行处理，以捕获多尺度特征. 前 3 个分支采用 3 个不同空洞率的空洞卷积对特征图进行卷积，再经过 ReLU 激活函数. 空洞卷积的空洞率为 r，卷积核大小由 $k\times k$ 扩大到 $k'=k+(k-1)\times(r-1)$. 因此，需要调整空洞卷积的空洞率，以适用于目标检测

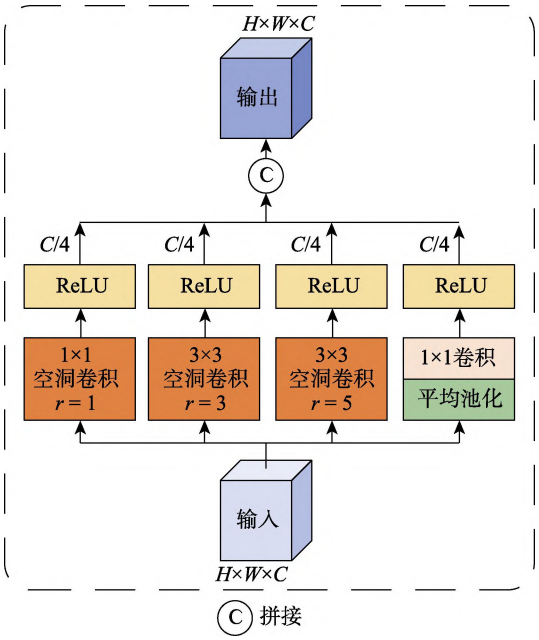


图 4 ASPPM 结构图

任务，设置卷积核大小 k=[1, 3, 3]、空洞率 r=[1, 3, 5]、填充 p=[0, 3, 5]、步长均为 1. 其中，卷积的输出通道数为输入通道数的 1/4. 第 4 个分支先经过平均池化操作压缩特征，再进行 1×1 卷积和 ReLU 激活函数，最后调整第 4 个分支输出特征的大小再与前 3 个空洞卷积分支的输出特征按通道维度拼接，得到输出结果. 在 ASPPM 中，应用 ReLU 激活函数引入非线性属性，使之具有更强大的拟合能力[13]，将 ASPPM 用于自顶向下路径中以丰富特征.

2.2.2 WFF

在 PAEM 中，利用 WFF 有效地融合 ASPPM 的输出特征与自底向上路径的输出特征，以进一步增强特征的表征能力. WFF 过程如下：f_i 为第 i 层级 ASPPM 的输出，f_i' 为第 i 层自底向上路径的输出. 首先，将 f_i 送入 1×1 卷积和 Sigmoid 激活函数计算得到注意力权重，它类似于软化的掩膜，用

于关注重要信息而忽略无关信息. 因此，选用 Sigmoid 激活函数将特征图归一化到 $(0,1)$，相当于获得特征图中每个像素值的关键程度权重；其次，将注意力权重分别作用于 f_i 和 f_i' 进行加权计算，并将其结果相加作为 WFF 输出结果，以达到关注特征图中感兴趣的区域的目的. 计算公式为

$$F_i=\sigma\left(C_{1\times1}\left(f_i\right)\right)\times f_i+\left(1-\sigma\left(C_{1\times1}\left(f_i\right)\right)\right)\times f_i' \tag{1}$$

其中，$C_{1\times1}(\cdot)$ 表示 1×1 卷积；$\sigma(\cdot)$ 表示 Sigmoid 激活函数；F_i 表示输出结果. 由于仅对 P_4, P_5 层级特征图运用 WFF，因此取 i =4, 5. WFF 结构图如图 5 所示.

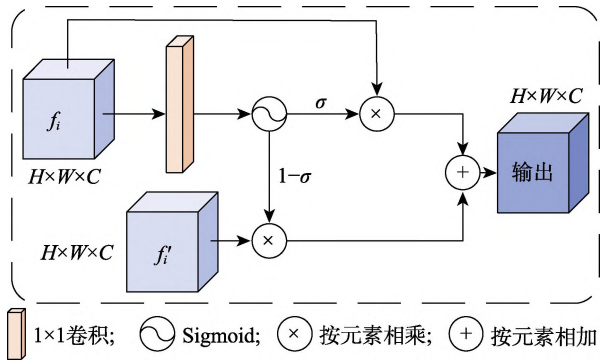


图 5　WFF 结构图

2.3　PBAM

CNN 提取到的低照度图像特征存在噪声干扰，无法生成显著的判别特征[9]. 金字塔结构能够提取不同尺度的特征图，但是因缺乏全局上下文信息导致无法在空间和通道维度对重要特征进行响应. 注意力机制能够突出重要信息而忽略无关信息，但无法有效地提取多尺度特征. 因此，本文结合多尺度特征表征和注意力机制，提出 PBAM. 该模块利用多尺度特征捕获模块(multi-scale feature capture module, MSFCM)学习不同尺度的特征空间信息形成金字塔结构，然后引入 BAM[25]生成包含不同尺度、更具判别力的特征，以减少噪声干扰的影响，进而提升模型的检测性能.

PBAM 结构图如图 6 所示，具体运算过程如下：首先，对输入特征图运用 MSFCM 以生成不同尺度特征图 f_i，将 f_i 按通道维度拼接生成特征图 F；其次，将 f_i 分别送入并联的空间注意力模块(spatial attention module, SAM)和通道注意力模块(channel attention module, CAM)计算权重，将得到的权重进行广播乘法得到部分注意力图 a_i，将 a_i 按通道维度拼接得到整体注意力图 A；然后，将 A 与 F 按元素相乘得到注意力加权后的特征图 F'；最后，将 F' 与输入特征图按元素相加得到输出特征图. 其中，BAM 中并行的 SAM 和 CAM 分别在空间和通道维度对特征图进行权重分配，对于重要的位置进行特征响应，便于网络突出重要信息而忽略无关信息. 跳跃连接解决了网络退化和梯度消失等问题. 经过 PBAM 生成具有不同尺度、更具判别力的特征. PBAM 公式表示为

$$\begin{cases} a_i=F_{\mathrm{SAM}}\left(f_i\right)F_{\mathrm{CAM}}\left(f_i\right),\ i=0,1,2,3 \\ A=\mathrm{Cat}\left[a_0,a_1,a_2,a_3\right] \\ O=\left(F\times A\right)+I \end{cases} \tag{2}$$

其中，f_i 表示经过 MSFCM 得到不同尺度的特征图；$F_{\mathrm{SAM}}(\cdot)$ 表示 SAM 操作；$F_{\mathrm{CAM}}(\cdot)$ 表示 CAM 操作；a_i 表示不同尺度特征图的部分注意力图；$\mathrm{Cat}[\cdot]$ 表示按通道维度拼接；A 表示整体注意力图；F 表示不同尺度的特征图 f_i 按通道维度拼接得到的多尺度特征图；I 表示输入特征图；O 表示输出特征图；• 表示广播乘法；× 表示按元素相乘；+ 表示按元素相加.

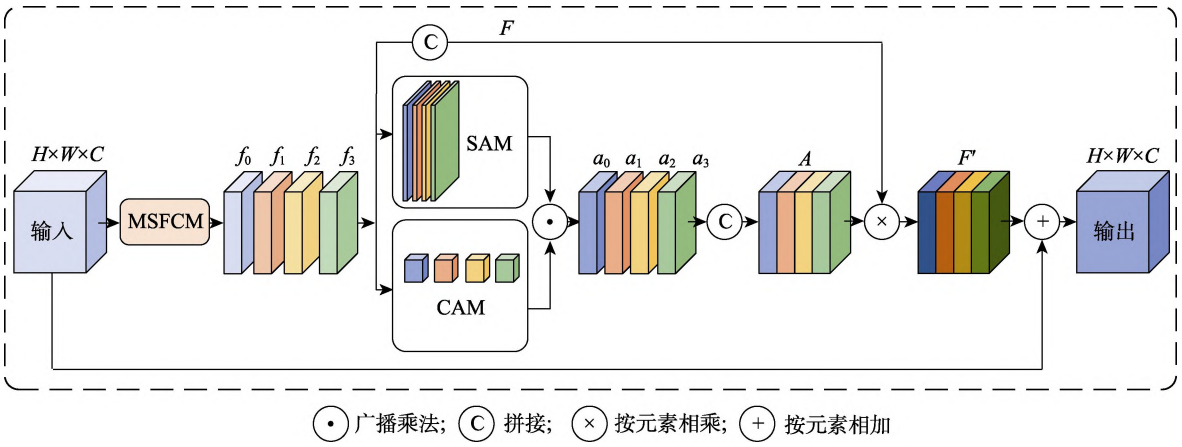


图 6　PBAM 结构图

2.3.1 MSFCM

MSFCM 以并行的多分支结构独立地学习输入特征图的空间信息，可以获取多尺度特征并形成金字塔结构，因此，其有助于经过并行的 SAM 和 CAM 标注生成更具判别力的特征图. 在金字塔结构中，使用具有不同卷积核大小的卷积操作可以获得不同尺度的特征空间信息. 为减少参数量，MSFCM 使用深度可分离卷积[28]而非普通卷积提取输入特征图的空间信息.

MSFCM 结构图如图 7 所示，输入特征图分别送入具有 4 个不同大小卷积核的深度可分离卷积. 设卷积核 $k_i=\{1, 3, 5, 7\}$，输出通道数均为输入通道数的 1/4，将得到的特征图按通道维度拼接得到输出特征图，以丰富输入特征图的特征空间. 多尺度特征图计算公式为

$$\begin{cases} f_i = C_{k_i \times k_i}(I), i = 0,1,2,3 \\ F = \mathrm{Cat}\left[f_0, f_1, f_2, f_3\right] \end{cases} \tag{3}$$

其中，$C_{k_i \times k_i}(\cdot)$ 表示第 i 个深度可分离卷积分支.

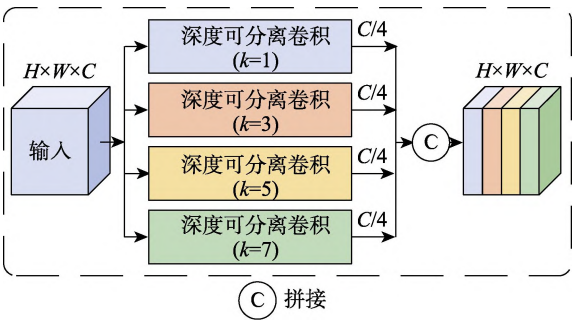


图 7 MSFCM 结构图

2.3.2 CAM

CAM 通过为输入特征图中越重要的通道赋予更大的权重，旨在让网络选择重要的通道进行特征响应. 首先，使用平均池化操作对空间维度进行特征压缩，以避免空间维度中噪点导致错误权重，得到 $1\times1\times C$ 的特征图. 其次，将该特征图送入由 2 个 1×1 卷积层和 1 个非线性激活函数 PReLU 组成的多层感知机(multi-layer perceptron, MLP)学习通道权重. PReLU 激活函数是在 LeakyReLU 的基础上改进的，其将负值部分的斜率 α 作为参数进行训练，在 CAM 中为增加 MLP 的非线性使用 PReLU 激活函数[25]. 为减少参数量和计算量，MLP 使用瓶颈结构，第 1 个卷积层压缩通道 4 倍，第 2 个卷积层还原通道数. 最后，Sigmoid 激活函数将结果归一化到 $(0,1)$，得到通道注意力权重

$$F_{\mathrm{CAM}}(I) = \sigma\left(F_{\mathrm{MLP}}\left(F_{\mathrm{AvgPool}}(I)\right)\right) \tag{4}$$

其中，I 表示输入特征图；$F_{\mathrm{AvgPool}}(\cdot)$ 表示平均池化操作；$F_{\mathrm{MLP}}(\cdot)$ 表示 MLP 操作；$\sigma(\cdot)$ 表示 Sigmoid 激活函数. CAM 结构图如图 8 所示.

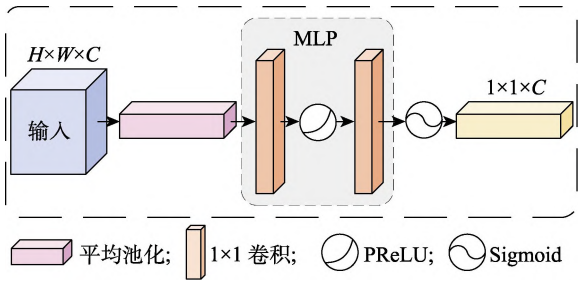


图 8 CAM 结构图

2.3.3 SAM

SAM 压缩通道维度为输入特征图计算空间注意力权重，旨在找到低照度图像中空间位置贡献更大的区域. 为保留通道维度中以最大值方式存在的高频细节信息，首先，用最大池化操作压缩通道维度，得到 $H\times W\times1$ 的特征图. 其次，对该特征图使用具有较大感受野的 7×7 卷积对其需要增强或抑制的空间位置信息进行学习. 最后，Sigmoid 激活函数将结果归一化到 $(0,1)$，得到空间注意力权重

$$F_{\mathrm{SAM}}(I) = \sigma\left(C_{7\times7}\left(F_{\mathrm{MaxPool}}(I)\right)\right) \tag{5}$$

其中，I 表示输入特征图；$F_{\mathrm{MaxPool}}(\cdot)$ 表示最大池化操作；$C_{7\times7}(\cdot)$ 表示 7×7 卷积；$\sigma(\cdot)$ 表示 Sigmoid 激活函数. SAM 结构图如图 9 所示.

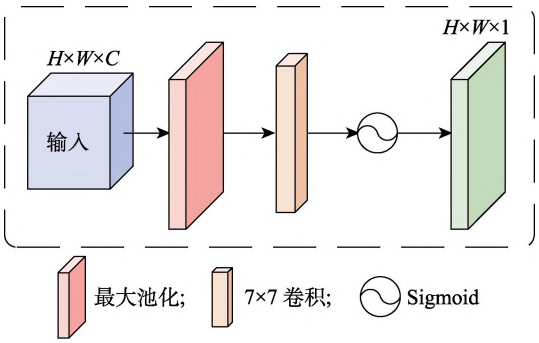


图 9 SAM 结构图

2.4 预测 IoU

在 YOLOv4 网络模型中，预测框的置信度是由分类得分和目标性得分相乘计算得到的，然而其未考虑预测框的定位精度. 因此，可能存在分类得分相对较高但定位精度低的预测框被保留，分类得分相对较低但定位精度高的预测框却被丢弃，

从而导致检测精度下降. 受感知 IoU[29]的启发，为了在低照度场景下能够更准确地定位目标，使用预测 IoU 改进 Head. 具体做法是：增加 1 个 IoU 预测分支，为每个预定义的预测框预测 IoU 值衡量定位的准确性. 使用交叉熵损失用于训练 IoU 预测分支，并且仅对包含目标的预测框即正样本进行损失计算，计算公式为

$$L_{\mathrm{IoU}}=-\frac{1}{N_{\mathrm{pos}}}\sum_{i\in\mathrm{pos}}^{N}\left(t_i\times\log\left(p_i\right)\right) \tag{6}$$

其中，t_i 表示预测框和对应的真实框之间的 IoU 值，即目标 IoU 值；p_i 表示网络为每个预测框预测得到的 IoU 值；N_{pos} 表示正样本数量.

在训练期间，分类损失、置信度损失、边界框回归损失与 YOLOv4 方法相同，IoU 预测分支与分类分支、目标置信度分支以及边界框回归分支一起训练，总损失函数计算公式为

$$L_{\mathrm{detection}}=L_{\mathrm{cls}}+L_{\mathrm{obj}}+L_{\mathrm{reg}}+L_{\mathrm{IoU}} \tag{7}$$

其中，L_{cls} 表示分类损失函数；L_{obj} 表示置信度损失函数；L_{reg} 表示边界框回归损失函数；L_{IoU} 表示预测 IoU 损失函数.

在测试时，预测框的分类得分 c_i、预测得到的 IoU 值 p_i 的 $(1-\alpha)$ 次方和目标性得分 o_i 的 α 次方三者相乘，结果作为预测框的置信度 $S_{\mathrm{detection}}$，即

$$S_{\mathrm{detection}}=c_i\times p_i^{(1-\alpha)}\times o_i^{\alpha} \tag{8}$$

其中，参数 α 控制两者对预测框的置信度的作用，设参数 α =0.4.

因此，预测框的置信度不仅考虑分类得分、目标性得分，同时考虑预测框的定位精度，通过 IoU 预测分支，可以将那些低照度场景下分类得分相近而定位精度更高的预测框保留下来，从而进一步提高模型的定位精度.

3　实验及结果分析

3.1　实验数据集及实验细节

本文在低照度目标检测数据集 ExDark[9]上进行实验，对 Dark-YOLO 方法进行评估. ExDark 数据集总共有 7363 幅图像，按 8 :2 的比例划分为训练验证集和测试集. 从训练验证集中划分 10%的图像作为验证集，因此，将数据集划分为训练集、验证集和测试集，它们的图像数分别有 5301 幅、589 幅和 1473 幅. 数据集总共标注 12 个预定义类别，包括自行车(Bicycle)、船(Boat)、瓶子(Bottle)、公交车(Bus)、汽车(Car)、猫(Cat)、椅子(Chair)、杯子(Cup)、狗(Dog)、摩托车(Motorbike)、人(People)和桌子(Table).

为了验证 Dark-YOLO 方法的有效性，实验与目前几种通用目标检测方法进行对比. 实验硬件环境为处理器 Intel(R) Xeon(R) CPU E5-2620，64 GB 内存，NVIDIA GeForce GTX 1080 Ti 图形处理器；软件系统为 Windows Server 2016 Datacenter 操作系统，实验在 Pytorch 深度学习框架使用 Python 编程语言实现.

Dark-YOLO 在训练和测试时，将图像统一调整到宽高均为 416. 由于 ExDark 数据集数量较小，直接训练神经网络的效果并不理想. 因此，Backbone 使用 ImageNet 预训练模型，再使用 ExDark 数据集进行训练，以使神经网络的收敛速度更快. 在训练时，设置 Batch size 为 4，最初训练加载预训练模型后冻结 Backbone 参数，设学习率为 0.001，权重衰减因子为 0.000 5，并选择固定步长衰减模式更新学习率训练网络 25 个 epochs；更新 Neck 和 Head 的参数使得网络能够较好地得到检测结果；后续训练解冻 Backbone 参数，设学习率为 0.000 1,权重衰减因子为 0.000 5，并选择固定步长衰减模式更新学习率训练网络 25 个 epochs，使得网络模型达到最佳的性能.

实验结果以平均精度(average precision, AP)、均值平均精度(mean average precision, mAP)和帧速(frames per second, FPS)作为客观评价指标，设其 IoU 阈值为 0.5. IoU 是指预测框和目标真实框之间的交并比，用于评价是否成功预测目标的位置，认为当 IoU>0.5 时，成功预测得到目标框的位置. AP 值同时考虑精确率和召回率，针对不同召回率下的精确率求取平均值，即 P-R 曲线下方的区域面积，用于衡量每个类别的检测精度. mAP 值为所有预定义类别的 AP 值的平均值，用于评价网络模型的整体检测性能. 它们的计算公式为

$$\begin{cases} P=\dfrac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{P}}} \\ R=\dfrac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{N}}} \\ \mathrm{AP}=\displaystyle\int_0^1 P(r)\mathrm{d}r \\ \mathrm{mAP}=\dfrac{\displaystyle\sum_{i=1}^{n}\mathrm{AP}(i)}{n} \end{cases} \tag{9}$$

其中，T_{P} 表示正确预测的正样本，即正确检测到

目标的数量；F_P 表示错误预测为正样本的负样本数量，即在检测目标时误检目标的数量；F_N 表示错误预测为负样本的正样本数量，即模型出现漏检目标的数量；P 表示精确率；R 表示召回率；$P(r)$ 表示不同召回率 r 下的精确率，即 P-R 曲线；n 表示预定义的类别总数；$\mathrm{AP}(i)$ 表示某一类别 i 的平均精度. 同时，检测速度也是衡量性能的重要指标，FPS 为每秒钟可以检测的图像数量，计算公式为

$$\mathrm{FPS}=\frac{N_{\mathrm{frames}}}{T_{\mathrm{time}}} \tag{10}$$

其中，N_{frames} 表示检测图像总数；T_{time} 表示检测总时间.

为了验证模型的 Batch size 大小设置为 4 时 Dark-YOLO 性能最佳，本文使用 Batch size 大小分别为 1, 2, 4 和 8 训练网络，对 Dark-YOLO 的 Batch size 定量分析实验结果如表 1 所示.

表 1 Batch size 定量分析实验结果

Batch size	mAP/%
1	55.01
2	70.61
4	**74.76**
8	73.99

注. 粗体数字表示最佳实验结果.

3.2 消融实验

为了验证 Dark-YOLO 方法所提出的 PAEM, PBAM 和预测 IoU 是否有利于提升低照度图像目标检测模型精度，对网络模型进行了客观评价指标对比，并对实验结果进行分析.

表 2 所示为 Dark-YOLO 在 ExDark 数据集上消融实验结果，展示以 YOLOv4 网络为基准和使用不同模块的组合模型在 ExDark 数据集上每个类别的 AP 值和 mAP 值. 为获得低照度图像中丰富的特征便于后续的分类和回归任务，Dark-YOLO 使用 CSPDarkNet-53 网络作为 Backbone 用于特征提取. 实验在 CSPDarkNet-53 网络的基础上分别使用 PAEM 记为 M_1, PBAM 记为 M_2, 使用 PAEM 与 PBAM 记为 M_1+M_2, 组合成 3 种不同的网络模型，每组消融实验均按实验细节实现.

从表 2 可知，与 YOLOv4 网络相比，Dark-YOLO 的 mAP 提升了 4.10%. 实验结果表明，其中的 PAEM 和 PBAM 对低照度目标检测模型性能的提升具有积极的影响. 与 YOLOv4 网络相比，PAEM 的 mAP 提升 1.22%. 通过 PBAM 捕获多尺度特征，

表 2 ExDark 数据集上的消融实验结果

类别	AP/%				
	YOLOv4	M_1	M_2	M_1+M_2	Dark-YOLO
Bicycle	73.67	73.61	75.89	74.09	**76.32**
Boat	68.70	69.26	67.03	72.06	**73.29**
Bottle	70.63	68.10	69.08	72.48	**73.13**
Bus	87.64	90.17	89.97	90.35	**91.54**
Car	78.43	77.43	79.16	78.39	**80.46**
Cat	69.33	74.68	72.63	**76.31**	75.34
Chair	64.66	66.74	65.95	65.89	**68.38**
Cup	62.29	67.73	67.17	66.14	**69.50**
Dog	75.75	73.37	77.34	78.81	**82.52**
Motorbike	74.14	74.80	75.18	**75.27**	73.11
People	71.95	73.68	74.02	74.98	**75.91**
Table	50.69	52.99	54.87	56.78	**57.64**
mAP/%	70.66	71.88	72.36	73.46	**74.76**

注. 粗体数字表示最佳实验结果.

在空间和通道维度上突出重要信息，可减少低照度图像存在噪声干扰的影响，从而生成更具判别力的特征. 与 YOLOv4 网络相比，PBAM 的 mAP 提升 1.70%. 通过 PAEM 和 PBAM 的共同作用，mAP 提升了 2.80%. 与使用 PAEM 和 PBAM 的网络相比，引入预测 IoU 后的 mAP 提升 1.30%，表明通过预测 IoU，能够在具有挑战性的低照度场景下提高模型的检测精度.

为进一步验证 PAEM 和 PBAM 对低照度目标检测模型性能的提升具有积极影响，在运用 PAEM 的基础上，将 Dark-YOLO 方法中所有 PBAM 替换为相关注意力方法进行对比：SENet[22](记为 SE), CBAM[23], Coordinate Attention[24](记为 CoordAtt), BAM[25]等 mAP 值对比如表 3 所示. 对比结果表明，相关注意力方法对于网络检测精度的提升均有积极影响，均可突出重点信息忽略无关信息，进而提高特征的判别表征能力. 与其他注意力方法相比，由于 PAEM 和 PBAM 两者的共同作用，使得 Dark-YOLO 方法在低照度场景下具有更好的性能，可以减少网络误检和漏检情况的发生，提高模型的检测精度.

3.3 在 ExDark 数据集上的实验结果分析

本节在划分好的 ExDark 数据集实验，用于评估 Dark-YOLO 方法的有效性，并与目前几种主流的通用目标检测方法(包括 Faster R-CNN[2], YOLOv3[6], YOLOv4[7], SSD[8], CenterNet[15], EfficientDet[18], YOLOF[20]等)进行客观评价指标以及可视化结果对比、分析. 表 4 所示为它们的 mAP 值对比结果，可以看出，Dark-YOLO 方法具有更高

表 3　PBAM 与 4 种相关注意力方法对比

YOLOv4	PAEM	SE[22]	CBAM[23]	CoordAtt[24]	BAM[25]	PBAM	mAP/%
√							70.66
√	√						71.88(+1.22)
√	√	√					71.98(+1.32)
√	√		√				72.55(+1.89)
√	√			√			72.64(+1.98)
√	√				√		72.11(+1.45)
√	√					√	**73.46(+2.80)**

注. √表示使用该网络或模块，粗体数字表示最佳实验结果.

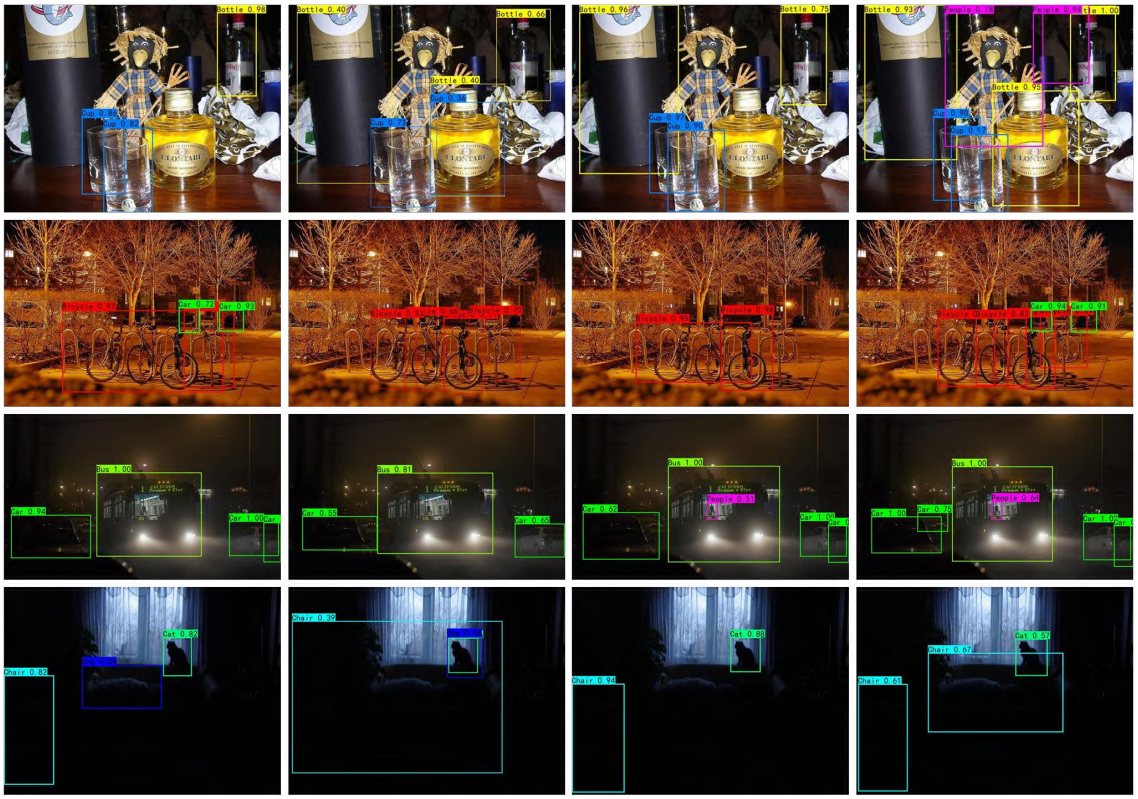
表 4　ExDark 数据集上不同方法的结果

方法	骨干网络	输入图像分辨率/像素	mAP/%
Faster R-CNN[2]	ResNet-50	800×800	63.52
SSD[8]	VGG-16	512×512	61.75
CenterNet[15]	ResNet-50	512×512	62.33
EfficientDet[18]	Efficient-B0	512×512	60.46
YOLOv3[6]	DarkNet-53	416×416	67.80
YOLOv4[7]	CSPDarkNet-53	416×416	70.66
YOLOF[20]	ResNet-50	800×1 333	68.05
Dark-YOLO	CSPDarkNet-53	416×416	**74.76**

注. 粗体数字表示最佳实验结果.

的精度. 与 YOLOv3[6]和 YOLOv4[7]方法相比，其 mAP 值分别提升了 6.96%和 4.10%. 网络的输入图像分辨率越大，则越有利于网络提取丰富的特征，进而提升网络的检测性能. 实验结果表明，与使用更大输入图像分辨率的目标检测方法 Faster R-CNN[2]，SSD[8]，CenterNet[15]，EfficientDet[18]，YOLOF[20]相比，Dark-YOLO 方法能够取得更好的检测结果，验证了其有效性和可行性.

Dark-YOLO，YOLOv3[6]，EfficientDet[18]和 YOLOv4[7]方法的可视化结果对比如图 10 所示. 可视化结果选择低照度下光照条件不一的图像，用检测框标注出目标类别和目标置信度. 可以看出，第 1 行中图 10a~图 10c 方法均存在漏检，由于背景区域较暗，这些方法均无法检测到人，只有 Dark-YOLO 方法能够成功地检测出所有目标；第 2 行中，由于目标与栏杆边缘细节特征定位不准确导致图 10a 和图 10c 方法均不能准确地得到自行车边界框，由于目标较小导致图 10b 和图 10c 方法漏检远处的汽车，Dark-YOLO 方法准确地检测到所



a. YOLOv3[6]　b. EfficientDet[18]　c. YOLOv4[7]　d. Dark-YOLO

图 10　在 ExDark 数据集上不同目标检测方法的可视化结果

有目标. 第 3 行中, 与图 10a~图 10c 所示方法相比, Dark-YOLO 方法能够检测到远处的小目标汽车; 由于右边 2 辆汽车重叠, 图 10b 所示方法不能有效地检测重叠场景的目标. 第 4 行中, 在人眼也无法辨认的情况下, 由于目标与背景区域区分度低, 图 10a~图 10c 所示方法均无法完全地检测到目标, 只有 Dark-YOLO 方法完全地检测到图像中的目标, 并准确地标注出边界框. 通过与其他方法可视化结果对比, Dark-YOLO 方法可以更好地完成低照度目标检测任务, 较好地解决了模型存在漏检和误检的问题, 提高模型的检测精度和目标定位的准确性.

表 5 所示为 Dark-YOLO 方法与不同目标检测方法在 ExDark 数据集上的 FPS 对比. 由于 PAEM, PBAM 和预测 IoU 会消耗一定的计算时间, 导致检测速度变慢, YOLOv4[7]和 YOLOF[20]方法的检测速度略快于 Dark-YOLO. 综合实验结果来看, 显然与 YOLOv3[6], YOLOv4, YOLOF 相比, Dark-YOLO 略微降低检测速度, 但显著地提升了检测精度, 并能够满足实时目标检测的要求. Dark-YOLO 方法对检测速度和检测精度有较好的平衡, 能够应用于对检测精度有一定要求的低照度场景.

表 5 ExDark 数据集上不同方法的 FPS 结果

方法	mAP/%	FPS
YOLOv3	67.80	**25.1**
YOLOv4	70.66	21.6
YOLOF	68.05	16.4
Dark-YOLO	**74.76**	15.3

注. 粗体数字表示最佳实验结果.

4 结 语

针对低照度场景下目标检测任务的检测精度较低、定位不准确等问题, 本文提出 Dark-YOLO 方法, 其使用 CSPDarkNet-53 网络作为 Backbone, 提出 PAEM 以增强特征的表征能力, 设计 PBAM 减少噪声干扰的影响, 还引入预测 IoU 改进 Head, 提高模型的定位准确性. 在低照度目标检测数据集 ExDark 上与主流的通用目标检测方法相比, Dark-YOLO 方法在低照度场景下具有更高的检测精度. 然而在检测速度方面有待提高. 下一步工作将研究轻量化的网络结构模型, 探索模块间的数据共享方法, 从而降低参数与运算量, 最终提高目标检测速度.

参考文献(References):

[1] Girshick R. Fast R-CNN[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 1440-1448

[2] Ren S Q, He K M, Girshick R B, *et al*. Faster R-CNN: towards real-time object detection with region proposal networks[C] //Proceedings of the International Conference on Neural Information Processing System. Cambridge: MIT Press, 2015: 91-99

[3] He K M, Gkioxari G, Dollár P, *et al*. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 42(2): 2980-2988

[4] 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. Los Alamitos: IEEE Computer Society Press, 2016: 779-788

[5] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 6517-6525

[6] Redmon J, Farhadi A. YOLOv3: an incremental improvement[OL]. [2021-09-10]. https://arxiv.org/abs/1804.02767

[7] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: optimal speed and accuracy of object detection[OL]. [2021-09-10]. https://arxiv.org/abs/2004.10934

[8] Liu W, Anguelov D, Erhan D, *et al*. SSD: single shot multibox detector[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2016: 21-37

[9] Loh Y P, Chan C S. Getting to know low-light images with the Exclusively Dark dataset[J]. Computer Vision and Image Understanding, 2019, 178: 30-42

[10] Sasagawa Y, Nagahara H. YOLO in the dark-domain adaptation method for merging multiple models[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2020: 345-359

[11] Chen C, Chen Q F, Xu J, *et al*. Learning to see in the dark[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 3291-3300

[12] 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. Los Alamitos: IEEE Computer Society Press, 2020: 390-391

[13] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[OL]. [2021-09-10]. https://arxiv.org/abs/1409.1556

[14] Lin T Y, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 2117-2125

[15] Duan K W, Bai S, Xie L X, *et al*. CenterNet: keypoint triplets for object detection[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 6568−6577

[16] He K M, Zhang X Y, Ren S Q, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition[C] //Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2014: 346-361

[17] Liu S, Qi L, Qin H F, *et al*. Path aggregation network for instance segmentation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 8759-8768

[18] Tan M X, Pang R M, Le Q V. EfficientDet: scalable and efficient object detection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 10778-10787

[19] Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks[C] //Proceedings of International Conference on Machine Learning. New York: PMLR, 2019: 6105-6114

[20] Chen Q, Wang Y M, Yang T M, *et al*. You only look one-level feature[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 13034-13043

[21] Mnih V, Heess N, Graves A. Recurrent models of visual attention[C] //Proceedings of the 27th International Conference on Neural Information Processing Systems. New York: ACM Press, 2014: 2204-2212

[22] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 7132-7141

[23] Woo S, Park J, Lee J Y, *et al*. CBAM: convolutional block attention module[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2018: 3-19

[24] Hou Q B, Zhou D Q, Feng J S. Coordinate attention for efficient mobile network design[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 13708-13717

[25] Wang F Y, Hu H T, Shen C. BAM: a lightweight and efficient balanced attention mechanism for single image super resolution[OL]. [2021−09−10]. https://arxiv.org/abs/2104.07566

[26] Huang Y S, Jiang Z T, Lan R S, *et al*. Infrared image super-resolution via transfer learning and PSRGAN[J]. IEEE Signal Processing Letters, 2021, 28: 982-986

[27] Chen L C, Papandreou G, Kokkinos I, *et al*. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848

[28] Howard A G, Zhu M L, Chen B, *et al*. MobileNets: efficient convolutional neural networks for mobile vision applications[OL]. [2021−09−10]. https://arxiv.org/abs/1704.04861

[29] Wu S K, Li X P, Wang X G. IoU-aware single-stage object detector for accurate localization[J]. Image and Vision Computing, 2020, 97: 103911-103932