第 38 卷　第 4 期
2023 年 8 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol. 38　No. 4
Aug. 2023

DOI:10.13364/j.issn.1672-6510.20220224

基于注意力与自适应特征融合机制的小目标检测

任克营，陈晓艳，茆　震，苗　霞，陈志辉
（天津科技大学电子信息与自动化学院，天津 300222）

摘　要：随着无人机平台的发展，航拍小目标检测成为当下研究热点。为了更有效地解决航拍小目标检测存在的漏检、错检以及重复检测等问题，提出了一种基于注意力与自适应特征融合机制的小目标检测算法 ST-YOLOX（small target-YOLOX）。本算法在 CSPDarknet 中融合了全局注意力模块（GC）以及可变形卷积（DC），增强主干网络对小目标特征的提取能力；采用四尺度自适应空间特征融合金字塔，抑制不同尺度之间的不一致信息，提升小目标特征表达的准确性；优化损失函数以及标签分配策略，提高算法检测精度。实验表明：ST-YOLOX 在 VisDrone-DET 2019 数据集中的平均检测精度（mAP）为 21.83%，比 YOLOX-s 模型提升了 3.78%，比 PPYOLOE-s 模型提升了 2.99%，比 YOLOv5-s 模型提升了 6.21%。航拍结果证明，本文算法的小目标检测准确率得到显著提高。

关键词：无人机航拍；单阶段检测算法；小目标检测；全局注意力机制；YOLOX；自适应特征融合
中图分类号：TP391.41　　**文献标志码**：A　　**文章编号**：1672-6510（2023）04-0054-08

Small-Target Detection Based on Attention and Adaptive Feature Fusion Mechanism

REN Keying，CHEN Xiaoyan，MAO Zhen，MIAO Xia，CHEN Zhihui
（College of Electronic Information and Automation，Tianjin University of Science & Technology，Tianjin 300222，China）

Abstract：With the development of UAV，drone-captured scenarios detection has become a hotspot of current research. In order to effectively solve the problem of missing，wrong and repeated detection caused by drone-captured scenarios detection，a novel algorithm named ST-YOLOX based on attention and adaptive feature fusion mechanism is proposed in this article. The algorithm combines the Global Context Module（GC）and Deformable Convolution（DC）in CSPDarknet to enhance the ability of backbone networks of extracting the features from small targets. A four-scale adaptive spatial feature fusion pyramid is used to filter the conflicting information between different scales and improve the expressive accuracy of the small target features. The loss function and label allocation strategies are applied to increase the target detection accuracy. Experiments showed that the mean average precision（mAP）of ST-YOLOX in the VisDrone-DET 2019 dataset reached 21.83%，which was 3.78% higher than that of YOLOX-s prototype，2.99% higher than that of PPYOLOE-s，and 6.21% higher than that of YOLOv5-s. Tests on the actual drone-captured scenarios verified that the accuracy of small-scale target detection was significantly improved.

Key words：drone shooting；one-stage detection algorithm；small target detection；global attention mechanism；YOLOX；adaptively spatial feature fusion

随着人工智能技术的不断发展，计算机视觉领域取得了巨大突破。目标检测作为计算机视觉的主要任务之一，目前已应用于行人检测[1-2]、人脸检测[3-5]等任务。主流算法有 SSD（single shot multibox detector）[6]、CornerNet[7]、YOLO（you only look once）[8-9]系列等，这些算法可以直接对目标进行分类和定位，无须生成大量的候选区域，因此有更快的检测速度，但其错误率以及漏检率也相对较高，尤其在无人机航拍

收稿日期：2022-09-29；**修回日期**：2022-11-12
基金项目：天津市科技支撑重点项目（18YFZCGX00360）
作者简介：任克营（1996—），男，天津人，硕士研究生；**通信作者**：陈晓艳，教授，cxywxr@tust.edu.cn

图像的小目标检测任务中。在 MS COCO[10]数据集上，小目标的检测精度甚至没有大/中目标检测效果的一半。因此，提高无人机航拍图像的小目标检测精度是亟待解决的问题。

2020 年，Nayan 等[11]提出一种基于 YOLOv3 的小目标检测算法，利用上采样和跳连接提取学习任务中不同卷积层的多尺度特征，显著提升了网络小目标检测能力。同年，郑晨斌等[12]提出一种强化上下文模块（enhanced context model，ECM），利用双空洞卷积结构减少参数量，扩大有效感受野，强化浅层上下文信息，并可以灵活应用于网络的浅层预测层。然而，这些方法严重依赖于上下文窗口的设计或感受野的大小，可能会导致重要上下文信息的丢失。2021 年，王建军等[13]提出一种改进 YOLOv3 的小目标检测算法，在主干网络中增加浅层特征对应卷积层网络的深度，以增强 backbone 对小目标特征的提取能力；引入 RFB（receptive field block）结构增大浅层特征图的感受野，提升小目标检测精度，在遥感图像小目标检测方面达到很好的效果。同年，旷视科技提出 YOLOX 算法，引入 Anchor free 思想，即在目标检测任务中无须设定预置锚框，但这使得目标检测中分类与目标定位任务之间缺乏交互，从而导致检测精度下降[14-15]。2021 年，基于自注意力机制的算法在许多视觉任务中取得显著效果，如 Swim Transformer[16]等。这类算法将自注意力机制应用到每个像素的局部窗口内，实现了比卷积神经网络（convolutional neural network，CNN）更好的效果，但其昂贵的内存访问成本导致检测明显比 CNN 慢。2022 年，百度科技提出 PP-YOLOE，该算法针对 Anchor free 算法类型的缺陷，引入了 TOOD（task-aligned one-stage object detection）中的 TAL（task alignment learning）[17]将目标检测中的分类与定位任务最优锚框拉近，进一步提升 YOLO 系列的检测精度。

为了实现高精度小目标检测方法，本文提出了一种新型的小目标检测算法 ST-YOLOX（small target-YOLOX）：

（1）提出一种基于全局上下文注意力主干提取网络 GD-CSPDarknet（global context deformable conv CSPDarknet）。

（2）采用四尺度自适应空间融合（adaptively spatial feature fusion，ASFF）方式抑制不同尺度之间的冲突信息。

（3）提出一种基于任务对齐策略的损失函数，优化标签分配方法。

1　ST-YOLOX 网络

YOLOX 与 ST-YOLOX 结构对比如图 1 所示，ST-YOLOX 主要包括骨干网络、Neck 网络、解耦检测头 3 部分，其基本网络结构遵循 YOLOX 的基本设计，即主干网络经过 3 次下采样。在浅层网络中，特征图分辨率较高，局部信息比较丰富，单个像素的感受野比较小，可以捕捉更多小目标的信息。因此，Neck 部分在原有基础上增加 1 次上采样操作，并增加了 1 个小目标检测头。

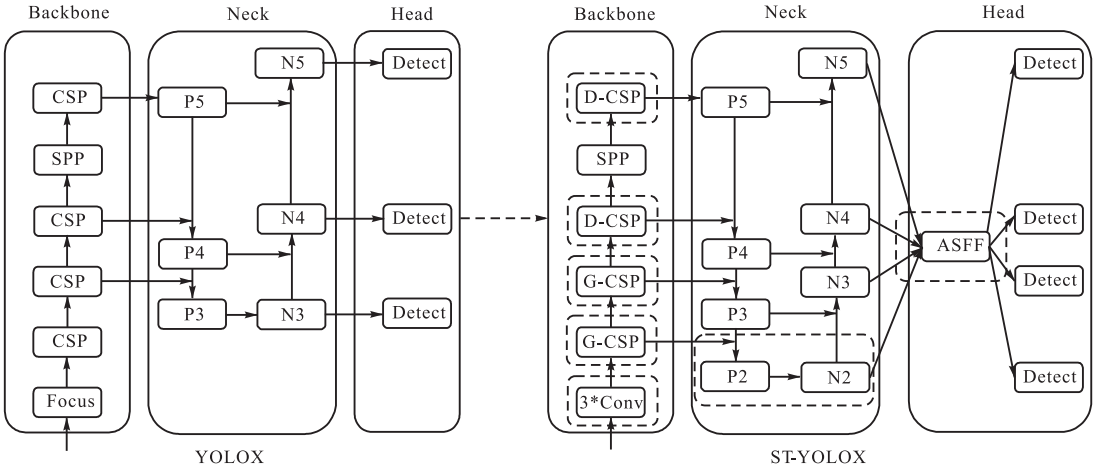


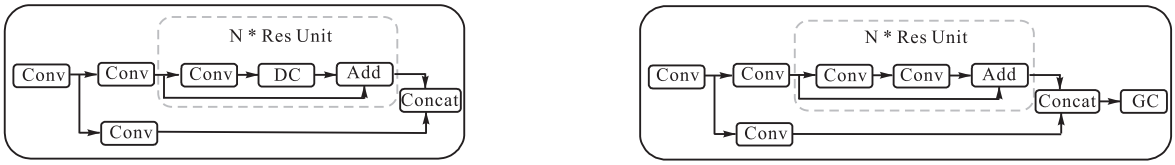
图 1　YOLOX 与 ST-YOLOX 结构对比

Fig. 1　Structure comparison of YOLOX and ST-YOLOX

1.1　骨干网络

传统 YOLO 算法骨干网络大部分采用残差连接[18-19]。残差连接通过使用身份映射的方式解决网络在训练过程中梯度消失的问题。CSP（cross stage

partial)[20]利用跳连接的方式降低计算负担，同时又尽可能减少模型精度损失。在小目标检测中，由于目标本身可利用的特征极少，这给检测任务增加很大难度，怎样利用全局信息提升小目标检测效果成为小目标检测研究的重点问题。本文提出了一种基于全局上下文注意力主干网络 GD-CSPDarknet。GD-CSPDarknet 能最大限度保持原有 CSPDarknet 架构的特点，采用 3 个堆叠的 3 × 3 卷积替换 Focus 模块，提升模型非线性拟合能力。在主干网络第一、二阶段的 concat 操作后引入全局上下文注意力模块(global context，GC)[21]，第三、四阶段用可变形卷积(deformable conv，DC)[22]替换原有的 3 × 3 卷积，其中 D-CSP、G-CSP 模块结构如图 2 所示。通过引入这两个模块将小目标及其周围上下文信息的特征最大化利用。



(a) D-CSP　　(b) G-CSP

图 2　D-CSP、G-CSP 模块结构

Fig. 2　Structures of D-CSP and G-CSP models

1.1.1　GC 模块

在早期特征中，目标语义信息较弱，且含有很多无用噪声，但早期特征信息有利于小目标定位，因此利用全局上下文注意力模块凸显对定位更有利的特征且抑制无用噪声，这成为早期特征处理的关键。GC 模块结构如图 3 所示。

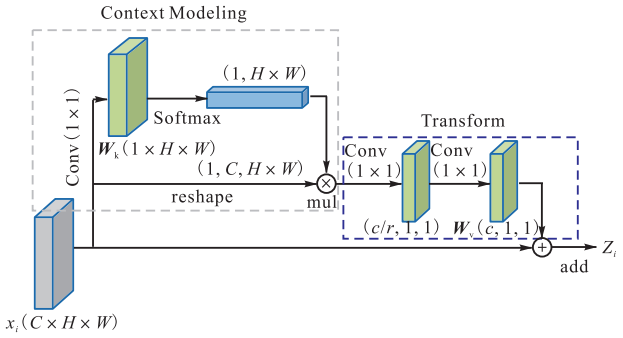


图 3　GC 模块结构

Fig. 3　Structure of GC model

该模块由两部分组成，第一部分为 Context Modeling，第二部分为 Transform。Context Modeling 为全局上下文注意力模块，该模块可以将特征图中相关联的特征聚合，形成全局上下文特征图。当该模块得到一个大小为 $C \times H \times W$(C 为通道数量，H 为图片长度，W 为图片宽度)的特征图之后，采用 1 × 1 卷积进行降维操作得到一个新的特征 $1 \times H \times W$，即为原始特征图的注意力权重 $\boldsymbol{W}_{\mathrm{k}}$，将 $\boldsymbol{W}_{\mathrm{k}}$ 的维度重塑为(1，$H \times W$)，再经过 Softmax 操作后得到一个权重分数(介于 0 ~ 1 之间)，将权重分数与重塑后的特征(1，C，$H \times W$)进行矩阵相乘，得到一个全局注意力热图，并将其维度重塑为(C，1，1)。Transform 模块采用 1 × 1 卷积进行通道压缩，再采用 1 × 1 卷积恢复到原始的特征维度。通过这种方式大幅减少网络参数数量。Transform 操作会得到一个新的权重矩阵 $\boldsymbol{W}_{\mathrm{v}}$，GC 模块计算公式为

$$Z_i = x_i + \boldsymbol{W}_{\mathrm{v}} \sum_{j=1}^{N_{\mathrm{p}}} \frac{\exp(\boldsymbol{W}_{\mathrm{k}} x_j)}{\sum_{m=1}^{N_{\mathrm{p}}} \exp(\boldsymbol{W}_{\mathrm{k}} x_m)} x_j \tag{1}$$

其中：x_i 为模块输入特征；Z_i 为模块的输出特征；N_{p} 为特征映射中的位置数，对于图像，$N_{\mathrm{p}} = H \times W$；$i$ 是查询位置的索引，j 列举出所有可能的位置。上式可以简化为如下 3 个操作：(1)通过 Context Modeling 生成一个全局注意矩阵，采用 1 × 1 卷积 $\boldsymbol{W}_{\mathrm{k}}$ 和 softmax 获得注意力权重，然后通过注意力权重获得全局注意力热图；(2)Transform 模块获得特征图中的通道相关性；(3)使用 Add 方式将全局上下文特征聚合到原始特征图中。

1.1.2　DC 模块

随着主干网络的加深以及下采样操作的进行，小目标的特征信息变得越来越弱。为了在深层主干网络中(第三、四阶段)尽可能利用全局信息加强小目标的表达能力，本文在深层主干网络中引入 DC，其结构如图 4 所示。

DC 模块是在卷积后加入向量偏置以及特征调制策略聚合小目标周围有利信息，其中向量偏置的特征维度为(1，$2k^2$，H，W)。$2k^2$ 为卷积核中所有元素在(x，y)方向的偏移量，即为图 4 的 offsets。特征调制标量 ΔM_k 可以调节来自输入特征中不同空间位置的

特征振幅，其取值在[0，1]范围内。在极端情况下，DC模块可以通过将特征幅值设置为零，决定不聚合该位置的特征。因此，来自相应空间位置的图像内容将显著减少，但不会对模块输出产生影响。其具体运算为

$$y(p)=\sum_{k=1}^{k} w_k \cdot x(p+p_k+\Delta p_k)\cdot \Delta M_k \tag{2}$$

其中 $y(p)$、$x(p)$ 分别表示输出特征映射 y 和输入特征映射 x 中位置 p 处的特征。

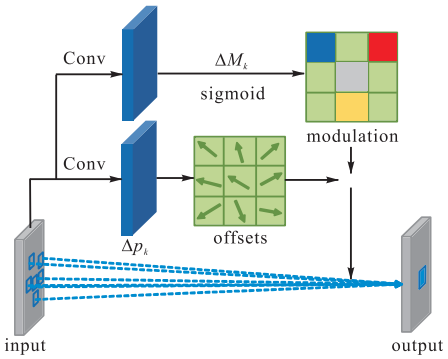


图 4　DC模块结构

Fig. 4　Structure of DC model

1.2　自适应空间特征融合

在目标检测算法中，为了适应不同尺度的目标检测，采用特征金字塔级联的方式进行特征融合。当使用特征金字塔检测对象时，大目标通常在深层特征中检测，小目标在浅层特征中检测，这会造成不同尺度特征之间的不一致性。因此，一幅图像中同时包含了小目标和大目标时，会在不同层次的特征之间产生冲突信息，这样就增加了小目标检测的难度。自适应空间特征融合（adaptively spatial feature fusion，ASFF）[23]可以抑制不同尺度特征之间在空间上冲突的信息。本文采用 ASFF 依次将 FPN + PAN 的 4 个输出 P_2、P_3、P_4、P_5 中 3 个相邻的输出进行空间及通道维度加权统一，权值 a_1^i、a_2^i、a_3^i 分别控制不同层级对融合信息的贡献。在每个空间位置上，对不同层次的特征进行自适应融合，即相同空间位置上的矛盾信息特征会赋予较小的权重，而对具有较高辨别性的特征赋予较大的权重，抑制不一致信息并加强有利信息的表征，通过训练找到最优的融合方式。融合后的特征 $y^i(x)$ 输出至解耦检测头。ASFF 的计算公式为

$$y^i(x)=\alpha_1^i P_n+\alpha_2^i P_{n+1}+\alpha_3^i P_{n+2} \tag{3}$$

其中：$i\in[1,4]$，$n\in[2,5]$。

1.3　解耦检测头

现有主流目标检测器仍采用耦合头实现目标检测中的目标分类及回归任务。本文将解耦检测头引入 ST-YOLOX 网络中，其结构如图 5 所示。该结构采用 1 × 1 卷积将传入预测头的特征通道压缩到 128，分别采用 2 个 3 × 3 卷积提取分类及回归两项任务所需要的特征，有效解决了分类与回归之间最佳锚点定位的冲突。

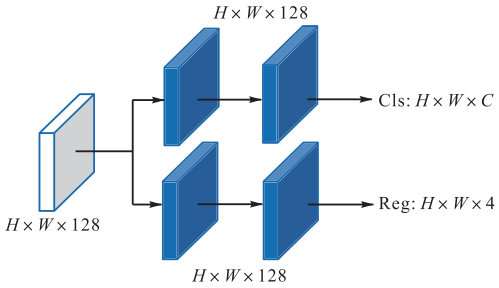


图 5　解耦检测头的结构

Fig. 5　Structure of decoupling detector

1.4　损失函数

针对分类任务的特点，本文引入了 VFL（varifocal loss）[24]作为损失函数。VFL 是一种动态尺度的二分类交叉熵损失函数，具体计算公式为

$$L_{\mathrm{VFL}(i,q)}=\begin{cases}-q(q\log i+(1-q)\log(1-i)) \\ -\lambda i^{\gamma}\log(1-i)\end{cases} \tag{4}$$

其中：i 为 IACS（IOU-aware classification score），即预测目标的联合 IOU（intersection of union）类别感知分数[24-25]；q 为目标 IOU 的得分。在训练过程中，如果该样本是正样本，则将 q 设置为预测框和真值框之间的 IOU，而对于训练中的负样本，所有类别的训练目标 q 均为 0。VFL 也会通过 i^{γ} 有效降低负样本损失的权重，正样本则不会降低权重。此外，通过 q 对 IOU 得分高的正样本损失加大权重，相当于将训练重点放在高质量的样本上。VFL 可以通过目标分数的排序衡量正样本的损失，这种排序对于 IOU 较高的正样本的损失贡献相对较大。这也使得模型在训练时更加关注高质量样本。通过 IACS 也可以有效地学习分类得分和定位质量估计的联合表示，从而实现网络在训练和推理之间的高度一致性。

对于回归任务，本文采用 DFL（distribution focal loss）[26]作为损失函数。DFL 可以使网络在任意灵活分布的条件下，快速专注于学习目标边界连续位置周围值的概率，提升特征边界不清晰目标的回归质量。具体计算公式为

$$L_{\mathrm{DFL}(S_i,S_{i+1})}=-((y_{i+1}-y)\log S_i+(y-y_i)\log S_{i+1}) \tag{5}$$

其中：$S_i=\dfrac{y_{i+1}-y}{y_{i+1}-y_i}$，$S_{i+1}=\dfrac{y-y_i}{y_{i+1}-y_i}$，$y_i$、$y_{i+1}$ 为标签 y

周围的数值。从式(2)可以看出，DFL 是回归一个离散域上的任意分布建模预测框，以类似交叉熵的形式优化与真值框最接近的左右两个位置的概率，从而让网络快速聚焦到目标位置邻近区域的分布中。

因此，兼顾分类和回归两项任务，ST-YOLOX 网络损失函数的最终表达式为

$$L=\frac{a\times L_{\mathrm{VFL}}+b\times L_{\mathrm{DFL}}+c\times L_{\mathrm{IOU}}}{\sum_{i}^{N_{\mathrm{pos}}}t} \tag{6}$$

其中 a、b、c 分别为 L_{VFL}、L_{DFL}、L_{IOU} 的权重。

1.5　标签分配策略

虽然优化损失函数可以拉近两个任务之间的差距，但仍未达到任务对齐的预期效果[17]。因此在 ST-YOLOX 模型中采用的标签分配策略为

$$t=s^{\theta}\times u^{\beta} \tag{7}$$

其中：s 为分类任务的得分，u 为预测框和真值框之间的 IOU。通过 θ 和 β 控制分类和 IOU 的得分对这个指标的影响程度。通过选择 m 个具有最大 t 值的锚框作为正样本，其余的为负样本。为了增加正样本锚框的得分，减少负样本锚框的得分，用 t 代替正样本锚框的标签质量。但是，当 θ 和 β 变换导致正样本的标签变小之后，模型将无法收敛，因此使用了归一化后的 t(即 $\hat{t}$)，可以在保持原来排序不变的前提下提升困难样本的学习效率。

2　实验及结果

2.1　数据集与评估指标

本文涉及的所有实验均在公开数据集 VisDrone-DET 2019[27]上进行。VisDrone-DET 2019 数据集由天津大学机器学习和数据挖掘实验室 AISKYEYE 团队收集。该数据集是无人机平台在不同地点、不同高度捕获的图像，共有 8 598 张图片，10 个类别，540 000 余个预标注锚框，其中训练集 6 470 张，验证集 548 张，测试集 1 580 张。

COCO 评价指标在目标检测中最为常用，同时也是最能反映一个目标检测算法性能好坏的指标。COCO 评价指标分为平均精确率(AP)、平均召回率(AR)两大类别。AP、AR 之后又细分为不同参数下的指标，其中 AP 划分为 mAP、AP50、AP75 等，AR 分为 AR1、AR10、AR100 等。mAP(mean average precision)是这些评价指标中最为常用也最能反映检测器性能的指标。具体计算公式为

$$R=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}}\times 100\% \tag{8}$$

$$P=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}}\times 100\% \tag{9}$$

$$P_{\mathrm{av}}=\int_{0}^{1}P(R)\mathrm{d}R \tag{10}$$

$$P_{\mathrm{avm}}=\frac{\sum_{i=1}^{k}P_{\mathrm{av},i}}{k} \tag{11}$$

其中：N_{TP} 表示正样本被预测为正样本的个数；N_{FP} 表示负样本被预测为正样本的个数；N_{FN} 表示负样本被预测为负样本的个数；R 为召回率(Recall)，P 为精确率(Precision)；P_{av} 表示以 Recall 值作为 X 轴、Precision 值作为 Y 轴的 $P(R)$ 曲线下覆盖的面积，该指标可以衡量某个类别的识别准确率；mAP 表示各类别 AP 的平均值，衡量在所有类别上的平均好坏程度，用符号 P_{avm} 表示。

2.2　实验设置

训练环境：Ubuntu 18.04 操作系统，CPU 为 Intel Xeon E5−2360，显卡为两块 NVIDIA GE−FORCE GTX1080Ti 11 GB，开发语言为 Python，深度学习框架为 PyTorch。

训练时使用随机梯度下降法(SGD)，参数设置：momentum = 0.9，weight decay = 5×10^{-4}；使用余弦退火学习率，初始学习率设置为零，预热训练 5 个轮次，基本学习速率为 0.01，通过线性缩放调整学习率。在训练过程中，采用指数移动平均(EMA)策略，其衰减为 0.999 8。在数据增强方面，去除了 YOLOX 中的数据增强策略，只保留了一些基本数据增强方法，如随机裁剪、随机水平翻转、颜色失真和多尺度训练，在以上参数设定下训练 300 个轮次。

2.3　消融实验

采用消融实验验证 ST-YOLOX 算法的有效性，以 YOLOX-s 为基线模型(baseline)，用 COCO 评价指标进行评价，结果见表 1。

表 1 中的第 1 行是以 YOLOX-s 为 baseline，复现了 YOLOX 原版工程，并遵循原版工程的所有设置。

表 1 中的第 2 行是在 baseline 的基础上优化了损失函数及标签分配的结果。模型的 mAP 从原来的 18.05% 提升到 18.49%，提升了 0.44%。这充分说明了优化损失函数及任务对齐策略后，检测效果有了明显的提升。

表 1 中的第 3 行是在前序基础上对特征提取骨干网络改进之后的结果。由此可以看出，优化特征提

取主干网络之后算法 mAP 又进一步提高到 18.98%，提升了 0.49%。这表明 GD-CSPDarknet 可以有效增强小目标特征提取能力以及小目标周围上下文信息聚合能力。

表 1 中第 4 行是在前序基础上提出四尺度 Neck 的结果。从表中可以明显看出，mAP 从原来的 18.98% 提升到 21.04%，提升了 2.06%。这表明增加小目标检测层能够进一步提升小目标的检测效果，这也印证了在 GD-CSPDarknet 中提出相关理论的合理性。

表 1 中的第 5 行指标是在之前优化模型的基础上进一步融合 ASFF 之后的指标。从数据上看，本模型的 mAP 从原来的 21.04% 提升到 21.83%，提升了 0.79%。这也充分说明了 ASFF 可以通过多尺度特征及融合操作过滤掉不同尺度之间特征的差异性。

表 1　消融实验结果对比

Tab. 1　Comparison of ablation experimental results

模型	mAP/%	AP50/%	AP75/%	AR1/%	AR10/%	AR100/%
YOLOX-s（baseline）	18.05	31.25	17.7	7.8	22.9	29.5
YOLOX-s + 改进损失函数	18.49	31.36	19.2	7.9	23.0	30.5
YOLOX-s + 改进损失函数 + GD-CSPDarknet	18.98	32.37	19.4	8.2	23.9	31.8
YOLOX-s + 改进损失函数 + GD-CSPDarknet + 四尺度 Neck	21.04	36.12	21.4	8.9	26.5	37.5
YOLOX-s + 改进损失函数 + GD-CSPDarknet + 四尺度 Neck + ASFF	21.83	37.32	22.5	9.0	27.3	38.5

综上所述，ST-YOLOX 比 YOLO-s 的 mAP 提升 3.78%，AP50 提升 6.07%，AP75 提升 4.8%，AR1 提升 1.2%，AR10 提升 4.4%，AR100 提升 9%。

2.4　对比实验

为了更进一步验证 ST-YOLOX 的检测性能，本文采用 YOLOX-s、YOLOv5-s、PPYOLOE-s 与 ST-YOLOX 进行检测性能对比，结果见表 2。测试图片的尺寸均为 640 像素 × 640 像素。

表 2　不同算法的检测结果

Tab. 2　Detection results of different algorithms

模型	图片尺寸	mAP/%	AP50/%
YOLOX-s	640 像素 × 640 像素	18.05	31.25
YOLOv5-s	640 像素 × 640 像素	15.62	28.74
PPYOLOE-s	640 像素 × 640 像素	18.86	32.21
ST-YOLOX	640 像素 × 640 像素	21.83	37.32

由表 2 可知：ST-YOLOX 比 YOLOv5-s 的 mAP 高 6.21%，AP 高 8.58%；比 PPYOLOE-s 的 mAP 高 2.97%，AP 高 5.11%。因此，ST-YOLOX 检测性能优于 YOLOX-s 等算法。

2.5　检测结果可视化

取实际场景航拍图像进行检测结果可视化，并与主流算法进行对比，结果如图 6 所示。图 6(a) 为待检测原版图片，将图片的红框标注部分进行放大，对该区域的小目标检测对比，图 6(b)—图 6(d) 分别为 YOLOX-s、YOLOv5-s、PPYOLOE-s 的检测结果。图 6(a) 中有 4 辆摩托车、2 辆三轮车、5 辆遮挡汽车。YOLOX-s 对密集摆放的摩托车只检出 1 辆，遮挡车辆只检出 2 辆，对电动三轮车造成误检，如图 6(b) 中红色圆圈标注。YOLOv5-s 对密集摆放的摩托车均没有检出，遮挡车辆只检出 3 辆，如图 6(c) 中红色圆圈标注。

(a) 原版图片

(b) YOLOX-s

(c) YOLOv5-s

(d) PPYOLOE-s

(e) ST-YOLOX

图 6　不同模型的检测结果可视化

Fig. 6　Visualization of test results of different models

PPYOLOE-s 中对密集摆放的摩托车虽然已经检出 3 辆，但是也对路边的自行车造成误检，遮挡车辆只检出 2 辆，如图 6(d)中红色圆圈标注。图 6(e)为本文所提出的 ST-YOLOX 算法的结果，可以看出相较于其余 3 种算法，密集摆放的摩托车以及遮挡车辆的检出率均有明显提升，没有出现误检的情况。

3　结　语

本文提出一种基于全局上下文注意力骨干网络 GD-CSPDarknet；采用四尺度自适应特征融合模块(ASFF)抑制不同尺度特征之间的冲突信息，从而改善不同尺度之间特征的不一致性；优化标签分配方法及损失函数，拉近目标检测任务中目标定位与分类任务之间的差距。实验结果表明，ST-YOLOX 算法在复杂场景下对小目标的检测效果较好，特别是在无人机航拍复杂城市场景下，优于当前主流目标检测算法。

虽然 ST-YOLOX 在性能指标方面有所提升，但将其应用于实际场景中仍存在一定的优化及改进空间，如优化在更复杂场景数据集上的检测效果以及将模型进一步轻量化，使模型更加有利于部署，实现边缘端的快速精准检测。

参考文献：

[1] DOLLAR P, WOJEK C, SCHIELE B, et al. Pedestrian detection: a benchmark[C]//IEEE. Proceedings of International Conference on Computer Vision & Pattern Recognition. New York: IEEE, 2009: 304–311.

[2] XU Y, DONG X, LIN S, et al. Detection of sudden pedestrian crossings for driving assistance systems[J]. IEEE Transactions on systems, man, and cybernetics, 2012, 42(3): 729–739.

[3] VIOLA P, JONES M J. Robust real-time face detection[J]. International journal of computer vision, 2004, 57(2): 137–154.

[4] JIE C, WANG R, YAN S, et al. Enhancing human face detection by resampling examples through manifolds[J]. IEEE Transactions on systems man & cybernetics part A systems & humans, 2007, 37(6): 1017–1028.

[5] CHEN G, HONG L, DONG J, et al. EDDD: event-based drowsiness driving detection through facial motion analysis with neuromorphic vision sensor[J]. IEEE Sensors journal, 2020, 20(11): 6170–6181.

[6] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//ECCV. Proceedings of European Conference on Computer Vision. Cham: Springer, 2016: 21–37.

[7] LAW H, DENG J. CornerNet: detecting objects as paired keypoints[C]//ECCV. Proceedings of European Conference on Computer Vision. Berlin: Springer, 2018: 765–781.

[8] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//IEEE. Proceedings of International Conference on Computer Vision & Pattern Recognition. New York: IEEE, 2016: 779–788.

[9] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//IEEE. Proceedings of IEEE Conference on Computer Vision & Pattern Recognition. New York: IEEE, 2017: 6517–6525.

[10] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//ECCV. European Conference on Computer Vision. Berlin: Springer International Publishing, 2014: 740–755.

[11] NAYAN A A, SAHA J, MOZUMDER A N. Real time detection of small objects[J]. International journal of innovative technology and exploring engineering, 2020, 29(5): 14070–14083.

[12] 郑晨斌，张勇，胡杭，等. 目标检测强化上下文模型[J]. 浙江大学学报(工学版)，2020，54(3)：529–539.

[13] 王建军，魏江，梅少辉，等. 面向遥感图像小目标检测的改进 YOLOv3 算法[J]. 计算机工程与应用，2021，57(20)：133–141.

[14] LI X, WANG W, WU L, et al. Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection[J]. Advances in neural information processing systems, 2020, 33: 21002–21012.

[15] YANG Y, LI M, MENG B, et al. Rethinking the aligned and misaligned features in one-stage object detec-

tion[EB/OL]. [2022-09-01]. https://arxiv.org/abs/2108.12176v1.

[16] LIU Z, LIN Y, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//IEEE. International Conference on Computer Vision. New York: IEEE, 2021: 1550-5499.

[17] FENG C, ZHONG Y, GAO Y, et al. Tood: task-aligned one-stage object detection[C]//IEEE. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). New York: IEEE, 2021: 3490-3499.

[18] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//IEEE. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016: 770-778.

[19] XIE S, GIRSHICK R, DOLLAR P, et al. Aggregated residual transformations for deep neural networks[C]// IEEE. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 1492-1500.

[20] WANG C Y, LIAO H Y M, WU Y H, et al. CSPNet: a new backbone that can enhance learning capability of CNN[C]//IEEE. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New York: IEEE, 2020: 390-391.

[21] CAO Y, XU J, LIN S, et al. GCNet: non-local networks meet squeeze-excitation networks and beyond[C]// IEEE. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. New York: IEEE, 2019: 1971-1980.

[22] ZHU X, HU H, LIN S, et al. Deformable convnets v2: more deformable, better results[C]//IEEE. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2019: 9308-9316.

[23] LIU S, HUANG D, WANG Y. Learning spatial fusion for single-shot object detection[EB/OL]. [2022-09-01]. http://www.arxiv-vanity.com/papers/1911.09516.

[24] ZHANG S, CHI C, YAO Y, et al. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection[C]//IEEE. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 9759-9768.

[25] JIANG B, LUO R, MAO J, et al. Acquisition of localization confidence for accurate object detection[C]//ECCV. Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018: 784-799.

[26] LI X, WANG W, HU X, et al. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection[C]//IEEE. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2021: 11632-11641.

[27] DU D, ZHU P, WEN L, et al. VisDrone-DET2019: the vision meets drone object detection in image challenge results[C]//IEEE. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. New York: IEEE, 2019: 213-226.

责任编辑：郎婧