引用格式：吴迪, 赵品懿, 甘升隆, 等. 基于动态自适应通道注意力特征融合的小目标检测 [J]. 电子科技大学学报, 2025, 54(2): 221-232.
WU D, ZHAO P Y, GAN S L, et al. Small object detection based on dynamic adaptive channel attention feature fusion[J]. Journal of University of Electronic Science and Technology of China, 2025, 54(2): 221-232.

基于动态自适应通道注意力特征融合的小目标检测

吴 迪[1,2*]，赵品懿[1]，甘升隆[1]，沈学军[1]，万 琴[1,2]

（1. 湖南工程学院 电气与信息工程学院，湘潭 411100；2. 湖南大学 机器人视觉感知与控制技术国家工程研究中心，长沙 410082）

摘要：针对小目标检测中卷积操作导致检测特征缺失和不同尺度语义隔阂的问题，提出一种基于动态自适应通道注意力特征融合的小目标检测方法。1）提出一种多尺度三角动态颈（Tri-Neck）网络结构，用于融合多尺度特征语义隔阂及弥补小目标特征缺失的问题。2）提出一种分组批量动态自适应通道注意力模块，增强弱语义小目标特征同时抑制无用信息，且在动态自适应通道注意力模块中设计新的激活函数和交并比损失函数，提升通道注意力表征能力。3）采用 ResNet50 作为骨干网络依次连接特征金字塔网络和 Tri-Neck 网络。实验结果表明，该方法在 Pascal Voc 2007、Pascal Voc 2012 上比 YOLOv8 算法 mAP 分别提升 5.3% 和 6.2%，在 MS COCO 2017 数据集上 AP 和 AP_S 分别提升 1.6% 和 2%，在 SODA-D 数据集上比 YOLOv8 算法 AP 提升 0.9%。

关键词：小目标检测；多尺度融合特征；特征金字塔；动态通道注意力；交并比损失函数
中图分类号：TP391.41; TH701 **文献标志码：**A DOI: 10.12178/1001-0548.2023270

Small object detection based on dynamic adaptive channel attention feature fusion

WU Di[1,2*], ZHAO Pinyi[1], GAN Shenglong[1], SHEN Xuejun[1], and WAN Qin[1,2]

(1. Institute of Electrical and Information Engineering, Hunan Institute of Engineering, Xiangtan 411100, China;
2. National Engineering Research Center of RVC, Hunan University, Changsha 410082, China)

Abstract: In order to solve the feature missing and different scales features semantics gaps problem caused by convolution operations in small object detection, a small object detection method based on dynamic adaptive channel attention feature fusion is proposed in this paper. Firstly, a Tri-Neck network structure is introduced to address the semantic gaps and feature deficiency in small object detection across multiple scales. Secondly, a dynamic adaptive channel attention module is proposed to enhance weak semantic features of small objects while suppressing irrelevant information. Additionally, new activation functions and intersection-over-union loss functions are designed within the dynamic adaptive channel attention module to improve channel attention representation capability. Finally, the ResNet50 backbone network is utilized, connecting the feature pyramid network and the Tri-Neck network sequentially. Experimental results on the Pascal VOC 2007 and Pascal VOC 2012 datasets demonstrate performance improvements of 5.3% and 6.2% respectively, while on the MS COCO 2017 dataset, the proposed algorithm shows enhancements in overall performance and small object detection performance by 1.6% and 2% respectively, and on the SODA-D dataset, our proposed algorithm demonstrates superior performance compared to the suboptimal algorithm AP, resulting in a 0.9% improvement in overall accuracy.

Key words: small object detection; multi-stage feature fusion; feature pyramid network; dynamic channel attention; iIOU loss

收稿日期：2023−10−31
基金项目：国家重点研发计划（2020YFB1713600）；国家自然科学基金（62476084）；湖南省教育厅重点项目（24A0528）；湖南省自然科学基金（2022JJ30198）；湖南省研究生科研创新项目（YC202213）
作者简介：吴迪，博士，副教授，主要从事多模态融合行人再识别和信息融合理论与应用等方面的研究。
*通信作者 E-mail: wudi6152007@163.com

在机器视觉领域中，目标检测是众多高级视觉的基础任务之一，小目标检测是目标检测中比较困难的任务。由于小目标图像特征在图像中的分布范围有限[1]，常被周围背景或其他目标所包围，对小目标检测造成了一定程度的干扰[2]。因此，如何在小目标数据稀缺下提升小目标检测性能和不同环境下检测方法的鲁棒性有重要的研究意义。

在计算机视觉领域，小目标是指在图像或视频中尺寸较小的目标物体。这些目标物体在图像中通常只占据一小部分像素，可用的纹理信息少、目标尺寸小，因此检测起来相对困难。小目标检测的主要任务就是在图像或视频中准确地检测出这些尺寸较小的目标物体，并为其绘制边界框。这些物体可能包括但不限于：医学影像中的细胞、病变区域等；交通监控中的车辆、行人等；无人机航拍中的建筑物、树木等；工业检测中的缺陷、瑕疵等。由于小目标的尺寸小、像素少，它们在图像中所占的面积通常较小，这使得它们容易被忽略或被误检。因此，小目标检测在计算机视觉领域具有挑战性。为了提高小目标的检测性能，研究者们提出了许多方法，包括使用更先进的神经网络模型，如 Faster R-CNN[3]、YOLOv2[4]、YOLOv3[5]、SSD[6] 等，这些模型在检测小目标方面有更好的性能；对网络进行针对性的优化，如通过调整网络结构、引入注意力机制、使用多尺度训练等方法，来提高模型对小目标的检测能力；使用集成学习方法，将多个模型的检测结果进行融合，提高检测小目标的准确性。

在特征表示方面，小目标的特征提取受限于其尺寸较小和通用特征提取范式。当前主流的特征提取器通常会对特征图进行下采样以减少空间冗余并学习高维特征，这不可避免地会降低小目标的表示能力。此外，小目标的特征往往会在卷积过程中受到背景和其他实例的干扰，使网络难以捕捉到关键的区分信息。为了解决这个问题，研究人员提出了一系列的方法，如 SE[7]（squeeze and excitation）、DyHead[8]（dynamic head）和 Res2net[9]。

基于卷积神经网络的目标检测方法可分为有锚框和无锚框两类。文献 [10] 提出 R-CNN（regions with CNN features）采用卷积神经网络提取候选区域中的特征，利用支持向量机对候选区域进行分类与回归。该算法拉开了有锚框目标检测算法的序幕，接下来 Faster R-CNN[3]、SSD[6]（single shot multibox detector）和 YOLOv2[4]、YOLOv3[5] 等系列工作均需要人工手动地设计锚框，但预定义锚框的大小和长宽比往往不适用于小目标的尺寸和形状分布。小目标的尺寸较小且多样性大，使用固定的锚框难以覆盖和捕捉到小目标的多样形状和尺寸。为提高检测的召回率，YOLOv2、YOLOv3 采用兴趣区域网络在图像上密集地放置锚框，因涉及预测框与真值框之间的交并比等复杂的计算，从而导致内存占用率高。为降低算法的计算复杂度，以 CornerNet[11]、CenterNet[12] 和 FCOS（fully convolutional one-stage object detection)[13] 为代表的无锚框的目标检测器被提出，其主要思想是不需要预定义锚框，使用特征点或边界框来表示目标的位置和大小，具有较好的适应性。但对于小目标而言，基于特征点或边界框进行位置和尺寸的预测，由于其尺寸较小且细节有限，因此预测的位置可能存在一定的定位误差，导致目标的定位不够准确。

无论是有锚框还是无锚框的检测器，不同尺度目标通过骨干网络下采样后，均会出现语义隔阂，导致小目标检测性能不佳。针对上述问题，文献 [14] 提出 Inception[15] 结构的变体网络以及增加卷积核感受野的空洞卷积[16]，利用金字塔池化[17] 操作增强特征提取能力，以降低语义隔阂。文献 [18] 提出特征金字塔网络（feature pyramid network, FPN），以解决因图像中物体尺度和语义复杂度变化导致特征图在不同层级上的特点互相矛盾，难以同时满足高分辨率和高语义信息需求的问题。文献 [19] 提出 CARAFE（content-aware reassembly of features），将特征上采样算子整合到 FPN 中以提升性能。文献 [20] 提出 AugFPN，通过一致监督、残差特征增强和软选择等方法对 FPN 进行改进。上述方法虽然利用特征融合来增强语义信息，一定程度上提升了检测性能，但相比于大中目标，仍然存在特征细节缺失或不同尺度的语义隔阂等问题，导致小目标检测效果依然较差。

基于此，本文提出一种基于动态自适应通道注意力特征融合的小目标检测方法。首先，本文提出一种多尺度三角动态颈（Tri-Neck）网络结构，利用 FPN 自上而下的多尺度特征平滑化连接和包含动态通道注意力自底而上的多尺度特征融合方式，同时捕捉不同尺度上的目标特征和小目标的关键信息，以抑制背景干扰，从而解决小目标过度抑制和特征畸变问题。其次，在 Tri-Neck 网络中，本文设计一种新的动态通道注意力模块，通过选择加权特征通道以提升模型表达和泛化能力，自适应关注任务关键通道，减少冗余特征，抑制背景表征，增强前景和背景之间的对比。最后，本文在动态通道注意力模块提出新的激活函数和交并比损失函数，

激活函数将注意力权重限制在 [0,1] 范围内，自适应控制当前路径动态阀的激活函数梯度变化，以增强有用信息通道并抑制冗余信息通道；交并比损失函数隐性考虑了预测框与真值框之间的高宽的损失，提高了预测框的匹配性。本文方法能够有效地融合具有丰富语义信息的低分辨率深层特征图与具有独特低维特征的高分辨率浅层特征图，显著提高小目标检测的准确性。

1　算法整体结构

本文算法首先采用 ResNet50[21] 作为骨干网络，自下而上地提取各尺度特征，利用特征金字塔结构自上而下融合具有高分辨率的浅层特征图和具有丰富语义信息的深层特征图。其次，将特征金字塔提取的多尺度特征输入到本文提出的自适应动态特征选择模块，以进一步提取细粒特征。本文设计的特征融合模块包括两种处理路径，同层级特征图的横向连接和相邻层级的下采样连接，本文在两种处理路径中均添加了动态通道注意力模块，以抑制各通道内的背景表征，从而提高前景和背景的对比度。同时，本文在动态通道注意力模块设计了新的激活函数和改进过的交并比损失函数。最后，对输出特征进行预测框回归和分类任务。本文算法的整体结构如图 1 所示。

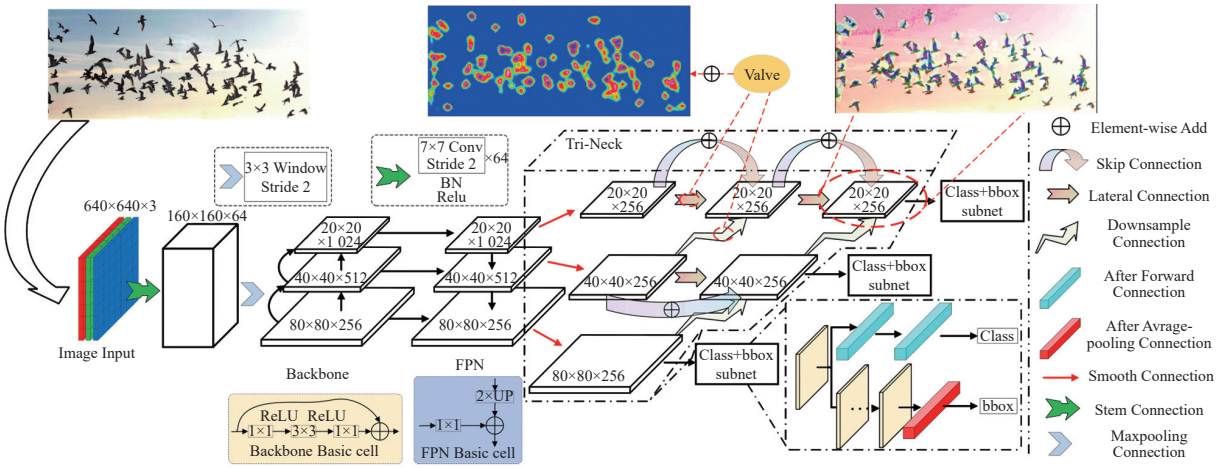


图 1　算法整体结构

2　特征提取模块

2.1　多尺度三角动态颈（Tri-Neck）

当前的目标检测通常采用卷积神经网络，通过逐层抽象提取目标特征。基于特征金字塔的方法旨在融合分辨率低、特征图感受野大且语义信息丰富的特征图，以及分辨率高、低维度特征明显（如几何信息）的特征图。然而，特征金字塔结构可能会导致过度抑制和特征失真的问题。由于不同尺度特征图中的信息差异过大，在分辨率低但语义丰富的特征图中，一些小目标的特征信息可能会被忽略，进而影响目标检测的精度。同时，在将较低分辨率的特征图上采样到较高分辨率时，可能会出现信息丢失和重采样等问题，导致特征失真。

基于此，本文提出一种 Tri-Neck 网络结构，如图 2 所示，该结构将经过平滑处理的特征图输入 Tri-Neck 网络中进行自下而上的融合，解决了特征失真的问题。图中 D1、P2'、P3'为经过 FPN 及平滑化连接得到的多尺度特征图。D1 分辨率高、语义信息缺乏但空间信息丰富；P3'分辨率低、语义信息丰富。在该结构中，对较高分辨率层的特征图进行下采样后，将其与相邻的下一层级的特征层融合。同时本文设计了两种融合的子路径，一种为横向连接，另一种为下采样连接，以更好地融合小目标相应尺度的特征。本文在各个连接中引入了动态注意力模块，使得小目标在高分辨率特征图中有更明显的几何特征，而在低分辨率特征图中有更明显的语义特征。

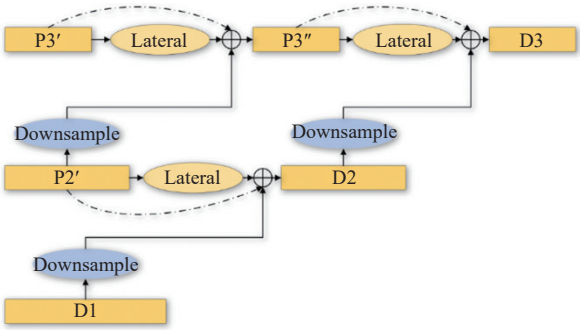


图 2　Tri-Neck 网络结构

网络层数某种程度上会影响网络结构性能，为进一步验证 Tri-Neck 网络结构的优越性，本文在

不同层数下验证本文方法的检测性能，具体如表 1 所示。当 Tri-Neck 层数为 4 的情况下，AP 性能达到 40.1%，性能得到了一定的提升，但进一步增加 Tri-Neck 层数参数，AP 性能提升不明显，通过增加层数的方式提升性能的效益明显降低。

表 1　使用不同尺度层数在 COCO2017 Val 数据集的 AP 性能对比

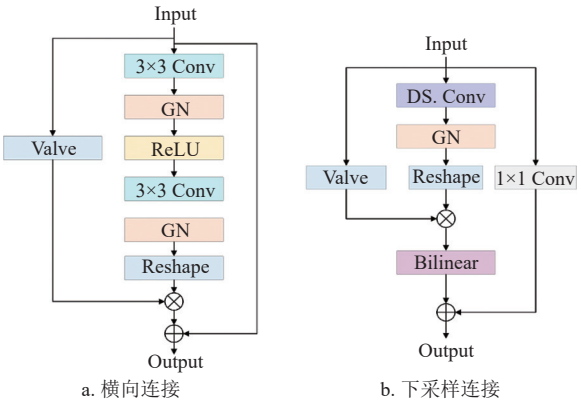
%

layer	AP	AP_{50}	AP_{75}
3	39	60.8	42.3
4	**40.1**	**64.2**	**44.7**
5	40.3	65.1	45.1

2.2　动态连接路径

本文在 Tri-Neck 网络结构基础上提出一种动态连接路径，一种是横向连接，另一种是下采样连接，从像素层面上缓解金字塔的缺陷。

横向连接路径经过深度可分离卷积加上组归一化激活后再经过一层深度可分离卷积[22]和组归一化[23]后与经过动态注意力权重相乘，最后加上残差连接得到输出，如图 3a 所示。下采样连接路径与横向连接相似，文献 [24] 认为现在的下采样操作不满足下采样定理，如最大池化、平均池化会引起输出生成剧烈波动。在使用最大池化的网络中，准确率并非随着偏移量的逐渐增大而一直下降，而是呈现周期性的震荡，这表明偏移量越大，网络效果不一定越差，因为周期性平移不变性在某些情况下仍然成立。当平移量为 N 的整数倍时，仍然可以满足平移不变性，使用双线性插值可以忽略这种影响，故本文直接使用双线性插值进行下采样，如图 3b 所示。



a. 横向连接　　b. 下采样连接

图 3　两种连接路径

在原始的自上而下的 FPN 结构中则采用的是邻近插值法，邻近插值法是一种更为简单的采样方法，它不会引入额外的参数，计算速度快，但采样后的结果可能会出现锯齿状的效果。因此，本文采用了双线性插值作为下采样的方案，相比于卷积类的下采样，双线性插值将原始图像中的每个像素点看作一个矩形区域，然后在下采样后的图像中，每个新像素点的位置处都有一个与原始图像中相应矩形区域大小和位置相同的矩形区域。最后，该新像素点的灰度值就由这个矩形区域内的像素点的灰度值插值计算得到。这种方法可以平滑地减小图像尺寸，同时保留图像中的细节信息。

动态增强颈部的特征图为$\{x_i^n\}^{C\times H\times W}$，其中$C$为通道数，$H\times W$为特征图大小，$n\in(1,2,3)$，$i\in(1,2,3)$。

$$X_{\mathrm{lc}}=v\left(x_i'^n,\delta_i^l\right)+x_i'^n \tag{1}$$

$$X_{\mathrm{ds}}=\mathrm{Bilinear}\left(v\left(x_i'^{n-1},\delta_i^{n-1}\right)\right) \tag{2}$$

$$\boldsymbol{F}=X_{\mathrm{lc}}+X_{\mathrm{ds}},\ \boldsymbol{F}\in\mathbb{R}^{C\times H\times W} \tag{3}$$

式中，$\boldsymbol{F}$表示骨干网络经过 FPN 对应层级的融合操作；$x_i'^n$表示经过处理后与动态注意力权重相乘的特征图。两种路径的内部都引入了动态注意力模块，X_{lc}代表横向连接的输出，X_{ds}代表下采样连接的输出，δ_i^l表示动态注意力中的权重。其中，动态注意力中权重δ严格限定在$\delta\in[0,1]$，本文提出了一个动态注意力的激活函数。

2.3　分组批量动态通道注意力

SE 模块仅通过建模通道之间的关系来重新权衡每个通道的重要性，但忽略了位置信息，而位置信息对于生成具有空间选择性的注意力图是重要的。在本文的两种路径中，为了更好地达到自适应动态效果，在动态混合中提出了分组批量动态通道注意力模块，既考虑了通道之间的关系又考虑了位置信息，如图 4 所示。

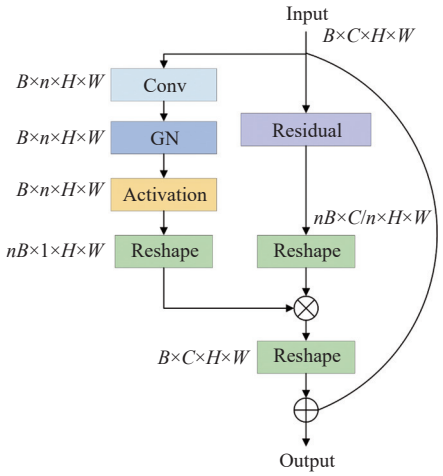


图 4　动态注意力机制

令输入特征图为$x^{B\times C\times H\times W}$，动态通道注意力权重为$\delta^{B\times n\times H\times W}$。$B$为批量大小，$C$为通道数，$n$为注

意力通道数量。

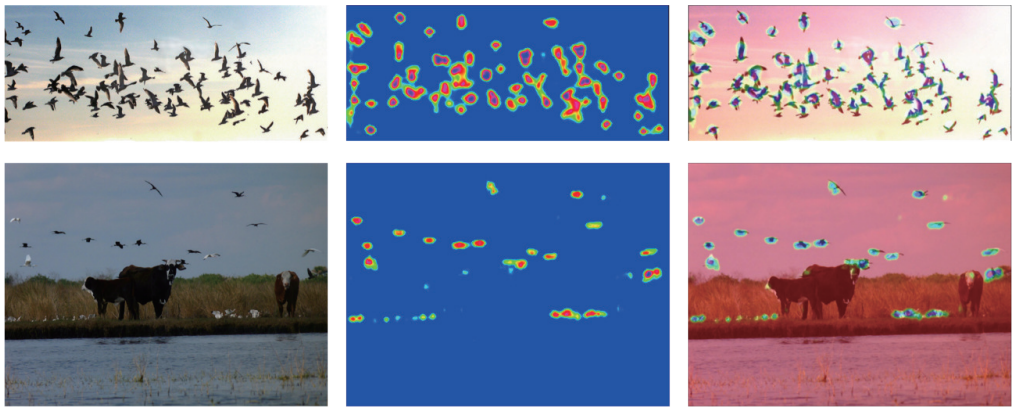
$$\delta = \sigma(\mathrm{GN}(\mathrm{Conv}(x))) \tag{4}$$

式中，σ为本文提出的激活函数；GN为组归一化操作。

最后，对输入的特征图x进行点乘加权：

$$\hat{x} = \delta x \tag{5}$$

动态通道注意力模块的优点在于它能够自动地学习和调整通道的重要性，提升网络对重要特征的感知能力，有助于模型更好地利用不同通道的信息，提高特征的表达能力和区分度。通过突出重要通道，模型能够更好地捕捉到数据中的关键信息，并抑制对无关信息的响应，提高模型的鲁棒性和泛化能力。本文保留高和宽上的维度以储存位置信息，通过保留位置信息和动态注意力模块，可以改善小目标在特征图上被忽略的问题并突出小目标位置。如图 5 所示，通过采用带有动态注意力 Tri-Neck 模型，本文增强了图像中小目标的语义信息和空间特征。通过观察小目标热力图，小目标区域显示出较高的注意力权重，表明模型对小目标有较高的关注度，意味着模型能够自适应地关注对小目标更有帮助的特征，从而提升感知和识别能力。同时，热力图中没有其他区域显示出高注意力权重，表明模型能够准确地聚焦在小目标上，避免了分散注意力。



a. 原始图　　b. 小目标热力图　　c. 叠加图

图 5　小目标热力图

3　激活函数和损失函数

3.1　动态注意力的激活函数

激活函数是动态注意力模块中很重要的一环，本文提出的动态注意力权重应严格限定在 [0,1]，用于动态地增强目标特征与抑制其他背景信息。为了更好地达到本文设计的预期，最后特征图背景像素数值经过激活函数作用后应该趋近于 0，另外为了获得更好的学习能力，在 0 点处应该是可导的。如果利用传统的 softmax[25]，当深度神经网络的层数较多时，容易出现梯度消失的问题，难以进行深层次的训练且不以 0 为中心，收敛速度慢。采用 Tanh 时，虽然以 0 为中心，但该激活函数的导数值域是 (0,1]，比 sigmoid 函数的 (0,0.25] 稍有缓解，但在输入值 x 的绝对值较大时（距离 0 较远时），导数仍然会趋近于 0，同样导致了梯度消失的问题。文献 [26] 提出 Tanh 激活函数用于选择不同动态路径点，但是限制性 Tanh 函数在 0 点具有不连续奇点，导致梯度在这一点上变化巨大。为了缓解这个问题，本文提出了一个更通用的激活函数，如式 (6) 所示：

$$\theta(v) = \max\left(\frac{1-\mathrm{e}^{-\alpha x}}{1+\mathrm{e}^{-\beta x}}\right) \quad \text{s.t.}[0,1], \quad \forall v \in \mathbb{R} \tag{6}$$

式中，α和β属于一个自适应动态调整的参数，α和β通过自适应学习，控制当前路径动态阀的激活函数梯度变化，该激活函数具有在零点连续且梯度缓和的特点。

3.2　交并比代价函数

文献 [27] 提出，当预测框与真值框不相交时，交并比（intersection over union, IOU）的值为 0。如果将其作为损失函数，那么它的梯度也是 0，将无法优化参数，也无法反映不相交的预测框与真值框之间的距离远近，因此不管距离远近，只要不相交，IOU 都为 0。为了解决这一情况，文献 [28] 提出了 GIOU，但是 GIOU 在实际应用中的效果并不理想，如果预测框与真实框不相交就退化成了 IOU。

$$L_{\mathrm{IOU}}=1-\left|\frac{A\cap A^{\mathrm{gt}}}{A\cup A^{\mathrm{gt}}}\right| \tag{7}$$

式中，A表示预测检测框；A^{gt}表示真实检测框。

基于此，本文提出一种新的 iIOU 函数，如图 6 所示，C 表示预测框和真实框的对角距离可以隐性地表达出两框之间的距离，角点之间的欧式距离约束可以反映出预测框和真实框中的长宽相似性。当两种框不相交时，IOU 等于 0，iIOU 就由对角点 C 来确定，如式 (8) 所示，当两种框距离越远，C 呈几何倍的增长，损失函数也就越大。当两种框相交时，IOU 大于 0，角点的约束使得两框之间长宽相似度逐渐升高，当两角点重合时退化为 IOU，如式 (9) 所示。当 IOU 等于 0 时，本文引入针对检测框尺度的损失函数，并增加了针对长宽的损失函数。这样的设计旨在提高目标框回归的稳定性，即使在目标与预测框完全不重叠的情况下，也能更好地调整框的尺寸和长宽比例。当 IOU 大于 0 时，本文考虑了每个检测框之间的欧氏距离。这个距离度量综合了目标与锚框之间的距离、重叠率以及尺度等因素。这种综合考虑旨在进一步提高目标框回归的稳定性，使得即使目标与预测框有一定重叠，也能更加准确地调整框的位置和尺寸，从而提高检测的精确度和稳定性。

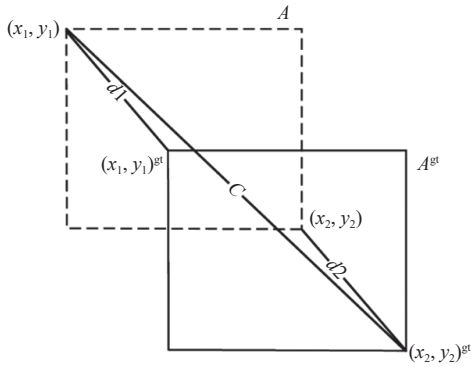


图 6 iIOU 示意图

当 IOU 等于 0 时：

$$L_{\mathrm{iIOU}}=1+\frac{c^2}{\rho^2((x_1,y_1)^{\mathrm{gt}},(x_2,y_2)^{\mathrm{gt}})} \tag{8}$$

当 IOU 大于 0 时：

$$L_{\mathrm{iIOU}}=1-\mathrm{IOU}+\frac{\rho^2((x_1,y_1),(x_1,y_1)^{\mathrm{gt}})}{c^2}+\frac{\rho^2((x_2,y_2),(x_2,y_2)^{\mathrm{gt}})}{c^2} \tag{9}$$

式中，(x_1,y_1)、$(x_1,y_1)^{\mathrm{gt}}$分别为预测框和真实框的左上角点坐标；(x_2,y_2)、$(x_2,y_2)^{\mathrm{gt}}$分别为预测框和真实框的右下角点坐标；ρ代表两点之间的欧式距离；c为(x_1,y_1)和$(x_2,y_2)^{\mathrm{gt}}$之间的欧式距离。

4 实验

本文模型训练与测试平台采用 Ubuntu 22.04 操作系统，Pytroch 2.0.0 深度学习框架，CUDA 11.8，CPU Intel i9-13900kf，内存 32 GB，显卡 GeForce RTX 4090 GPU（显存大小为 24 GB）。

4.1 训练细节

模型使用 AdamW 优化器进行优化，初始学习率设置为 2×10^{-4}，权重衰减设置为 0.05。批量大小设置为 8。使用 ResNet50 作为骨干网络提取多尺度特征，并使用 ImageNet[29] 上经过预训练的权重作为初始加载权重。在训练阶段，训练热启动设置为 5 个周期并采用线性热启动。训练图片经过随机水平翻转，归一化，Resize 大小为 1 333×800。测试图片 Resize 为 1 333×800。实验发现，当算法训练到 4 个 epoch 的时候趋于平稳，因此将学习率设置为初始学习率的十分之一，以实现更好的收敛效果。图 7 为训练过程中的收敛曲线，从图中可以看出在训练达到 12 个 epoch 后损失函数的收敛曲线不再下降。

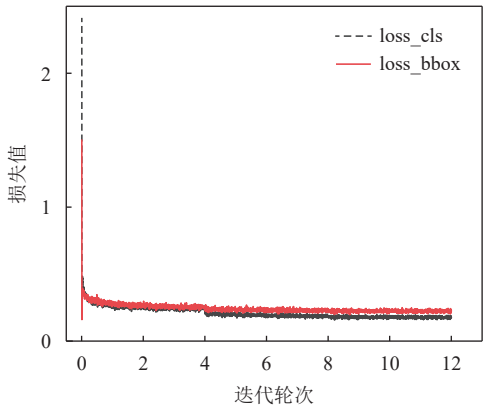


图 7 损失值与训练轮次的关系

4.2 数据集介绍

本文使用 MS COCO 2017、Pascal Voc 2007、Pascal Voc 2012 和 SODA-D 数据集进行实验。MS COCO 2017[30] 是微软公司发起的一个大规模通用物体检测、分割和图像理解数据集，包含超过 33.1 万张图片，其中超过 16 万张带有物体实例的标注，该数据集提供多种评估指标，包括 mAP、漏检率等。训练数据集采用 MS COCO 2017，数据集中划分出的训练集的图片为 118 287 张，该数据集中的物体实例包括 80 个常见类别，每个实例都

标注了其类别、边界框位置和图像分割掩码。Pascal Voc 是经典的目标检测和图像分割数据集，包含 20 个不同的物体类别，训练集和测试集各约 10 000 张图像。每个图像都有一个 XML 文件，包含该图像中每个目标的位置和类别标签。SODA-D 主要关注驾驶场景，包括 24 828 张高质量的图像和分布在 9 个类别中的 278 433 个实例：人员、骑行者、自行车、摩托车、车辆、交通标志、交通灯、交通摄像头和警示锥。SODA-D 最显著的优势之一是其多样性，包括时间跨度、地理位置、天气条件、摄像头视角等。

4.3　评价指标

为了更好更直观地与其他算法相对比，MS COCO 2017 采用所有类别平均精确度（AP）、小目标平均精确度的均值（APs）；Pascal voc 采用所有类别平均精确度的均值（mAP）、所有类别小目标平均精确度的均值（mAPs）作为本次实验的评价指标。在评价指标中，单个类别的 AP 是通过计算精确度−召回率（precision-recall）曲线下的面积来得到的。对于每个类别，根据不同的置信度阈值，计算在不同召回率下的精确度，并在整个召回率范围内进行插值。然后，对精确度−召回率曲线下的面积进行平均。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

$$\text{AP} = \int_0^1 \text{Precision d Recall} \tag{12}$$

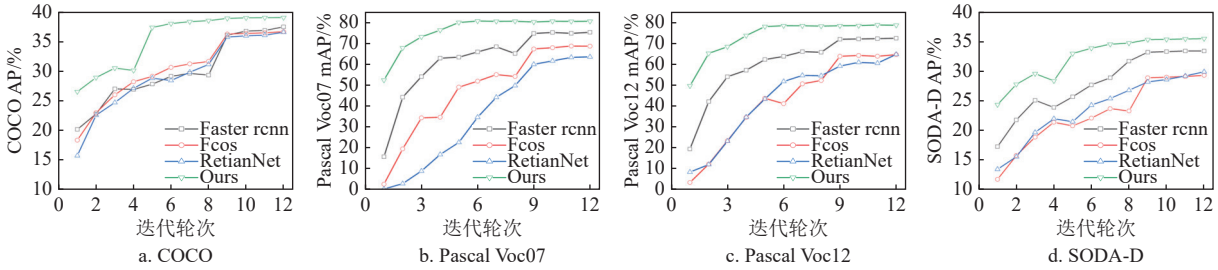
$$\text{mAP} = \frac{\sum_{i=1}^{C} \text{AP}_i}{C} \tag{13}$$

式中，TP 表示被正确地分类的正样本；FP 表示被错误分类的正样本；FN 表示被错误分类的负样本；C 表示当前检测的所有类别数量。

4.4　实验结果分析与比较

4.4.1　定量分析

为了测试本文算法的综合检测性能和小目标检测性能，本文选取 MS COCO 2017 中 Val 数据集、Pascal Voc 2007 中 test 数据集、Pascal Voc 2012 中 Val 数据集和 SODA-D 数据集作为测试集进行实验，将该模型算法与目前常用的几个算法进行比较，包括两阶段检测器 Faster R-CNN 一阶段检测器 FCOS、YOlOv3、RetinaNet[31]、YOLOX[32] 和 YOLOV8，实验结果如图 8 和表 2~表 5 所示。



a. COCO　b. Pascal Voc07　c. Pascal Voc12　d. SODA-D

图 8　4 种不同数据集下 AP 与 epoch 关系

表 2　不同算法在 COCO 2017 Val 数据集的 AP 性能对比

Model	backbone	epoch	AP/%	AP_{50}/%	AP_{75}/%	AP_S/%	AP_M/%	AP_L/%
Faster R-CNN-FPN	ResNet50	12	37.4	58.1	40.5	21.2	41.0	48.1
FCOS[13]	ResNet50	12	36.6	56.0	38.8	21.0	40.6	47.0
RetinaNet[31]	ResNet50	12	36.5	55.4	39.1	20.4	40.3	48.1
YOLOv3[10]	DarkNet53	300	30.8	52.8	32.0	14.4	33.4	44.7
YOLOX[32]	DarkNet53	300	31.8	49.1	33.8	12.4	34.9	47.4
YOLOv8n*	DarkNet53	300	32.8	46.9	—	—	—	—
Ours	ResNet50	12	**40.1**	**64.2**	**44.7**	**24.2**	**43.7**	**51.1**

注：Faster R-CNN原算法结构不含FPN，YOLOv8使用的评定标准是官方评定标准

表 3 不同算法在 Pascal voc 2007 test 数据集 AP 性能对比

Model	backbone	epoch	areo/%	bike/%	bird/%	boat/%	bottle/%	car/%	cat/%	chair/%	mAP/%
Faster R-CNN-FPN	ResNet50	12	82.6	82.7	77.7	61.4	63.5	84.9	85.7	57.3	75.2
FCOS[13]	ResNet50	12	75.9	73.0	72.8	55.2	60.8	81.6	83.0	55.6	68.6
RetinaNet[31]	ResNet50	12	65.9	75.5	72.3	46.1	59.9	82.3	77.2	52.9	63.6
YOLOv3[5]	DarkNet53	300	78.0	79.5	65.3	52.3	52.7	79.2	77.6	52.4	68.3
YOLOX[32]	DarkNet53	300	66.9	75.7	54.4	50.9	36.2	78.5	69.2	48.5	63.7
YOLOv8n*	DarkNet53	300	54.5	49.0	56.6	33.4	33.0	69.3	67.0	36.5	53.5
YOLOv8l*	DarkNet53	300	70.1	70.2	64.2	35.7	44.0	77.4	77.6	56.3	66.2
Ours	ResNet50	12	**86.4**	**86.7**	**80.0**	**71.6**	**69.9**	**88.2**	**89.3**	**63.4**	**80.5**

表 4 不同算法在 Pascal voc 2012 Val 数据集 AP 性能对比

Model	backbone	Epoch	areo/%	bike/%	bird/%	boat/%	bottle/%	car/%	cat/%	chair/%	mAP/%
Faster R-CNN-FPN	ResNet50	12	84.4	79.5	78.2	56.5	56.3	75.9	91.1	53.7	72.6
FCOS[13]	ResNet50	12	82.6	71.6	73.3	48.2	50.3	75.3	85.6	48.0	64.7
RetinaNet[31]	ResNet50	12	75.8	73.2	72.8	43.6	53.9	69.5	85.3	46.4	64.8
YOLOv3[5]	DarkNet53	300	78.1	74.2	63.1	44.8	48.1	65.7	83.5	48.0	65.3
YOLOX[32]	DarkNet53	300	75.9	70.0	56.3	43.4	35.7	67.1	77.3	43.3	61.9
YOLOv8n*	DarkNet53	300	59.7	55.3	55.4	39.5	40.2	75.3	69.2	40.3	57.6
YOLOv8l*	DarkNet53	300	72.5	73.1	65.2	41.4	42.5	74.8	76.5	62.6	69.1
Ours	ResNet50	12	**86.3**	**85.4**	**84.5**	**65.6**	**65.7**	**82.1**	**94.8**	**60.2**	**78.8**

表 5 不同算法在 SODA-D 数据集 AP 性能对比

Model	backbone	Epoch	AP/%	AP_{50}/%	AP_{75}/%	AP_T/%	AP_{eT}/%	AP_{rT}/%	AP_{gT}/%	AP_S/%	Param/M	FLOPs/G
Faster R-CNN[3]	ResNet50	12	33.4	64.3	29.9	29.3	19.7	30.5	36.7	43.8	41.16	292.28
FCOS[13]	ResNet50	12	29.3	55.7	26.5	24.6	11.4	25.7	33.2	41.3	70.75	137.21
RetinaNet[31]	ResNet50	12	29.9	58.7	25.5	25.3	15.9	27.2	32.3	40.3	35.68	299.50
RepPoints[33]	ResNet50	12	32.9	61.3	31.2	28.7	17.2	29.9	37.2	45.7	36.60	273.96
ATSS[34]	ResNet50	12	31.2	59.8	27.3	26.9	17.3	27.5	33.4	41.1	31.32	290.79
Deformable-DETR[35]	ResNet50	50	23.7	51.3	18.6	19.7	10.4	20.6	27.3	34.7	35.17	739.11
Sparse RCNN[36]	ResNet50	12	29.2	56.1	25.9	24.6	14.3	26.1	32.4	40.3	105.96	213.00
RFLA[37]	ResNet50	12	34.5	64.9	30.9	30.1	20.2	31.5	**38.5**	**46.8**	41.16	292.06
YOLOv8l*	DarkNet53	12	33.5	62.7	—	—	—	—	—	—	43.70	165.20
Ours	ResNet50	12	**35.4**	**65.3**	**31.4**	**30.7**	**20.4**	**32.2**	36.8	44.2	43.12	293.79

注：YOLOv8使用的评定标准是官方评定标准

图 8 展示了 4 种不同数据集下不同算法的 AP 与 epoch 关系。从实验结果可以看出，本文算法不仅最终综合性能优于其他算法，还加速了算法收敛速度。1）在收敛速度方面，Faster R-CNN 等算法在第 9 个 epoch 后才趋于收敛，而本文算法在第 6 个 epoch 已经趋于收敛，说明本文算法能够加速训练收敛速度。2）本文算法在第 1 个 epoch 时训练出来的性能可以超越其他算法在第 1 个 epoch 性能 5% 以上，在 Pascal Voc 2007 上甚至有 32% 性能提升，说明了该算法可能在复杂的任务或数据集中，能够更好地学习到具有较强判别能力的特征表示。

COCO 数据集在目标检测任务中使用了更严格的评估指标，如平均精确度（average precision, AP）和平均召回率（average recall, AR），以不同 IoU（交并比）阈值下的结果作为评估标准。COCO 数据集中的目标通常较小，因此对小目标的检测和定位提出了更高的要求。Pascal VOC 数据集使用平均精确度（average precision, AP）作为主要的评估指标，采用一组固定的 IoU 阈值进行评估。Pascal VOC 数据集中的目标通常较大和明显，因此对目标的定位精度要求相对较低。如表 2-表 5 所示，本文提出的基于动态自适应通道注意力特征融合的小目标检测在 MS COCO 2017、SODA-D[38]、Pascal Voc 2007 及 Pascal Voc 2012 中达到最优结果，分别超过次优结果 1.6%，0.9%，5.3%，6.2%。其中在 COCO 数据集小目标的 APs 上超越了次优模型 3% 以上。在 Pascal Voc 2007 和 Pascal

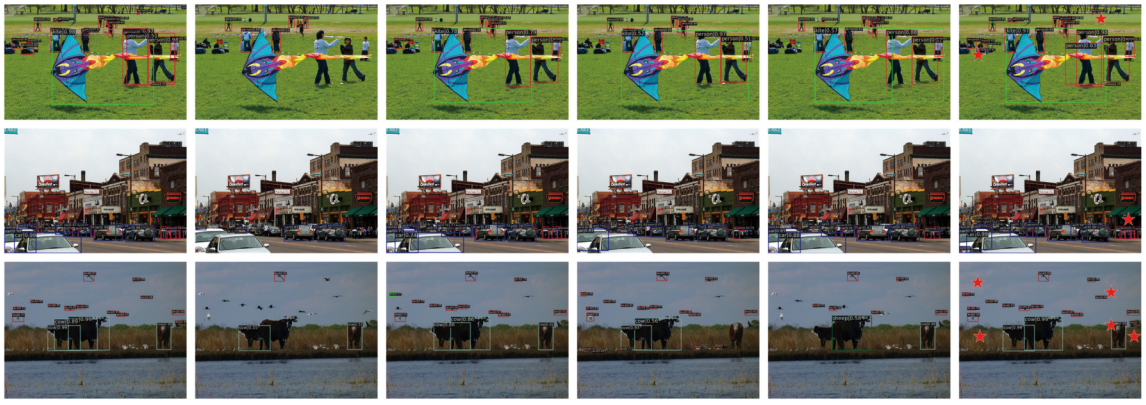
Voc 2012 数据集上，本文截取了 20 类中部分类别，其中对其他模型检测不佳的船类 AP 分别超过了次优结果 10.2% 和 9.9%，而对于鸟类的小目标检测性能领先次优模型 2.3% 和 6.3%。验证了本文提出的模型对小目标检测的有效性。

4.4.2　定性分析

为了进一步验证本文所提算法的有效性，本文在 MS COCO 2017 数据集上测试算法的检测性能。首先测试不同场景下，本文算法的检测性能，具体如图 9 所示，包括室内复杂场景、室外复杂场景、室内单一场景、室外单一场景、光线充足和光线不足下的场景。测试结果表明本文算法在不同场景下都有着不错的检测效果。

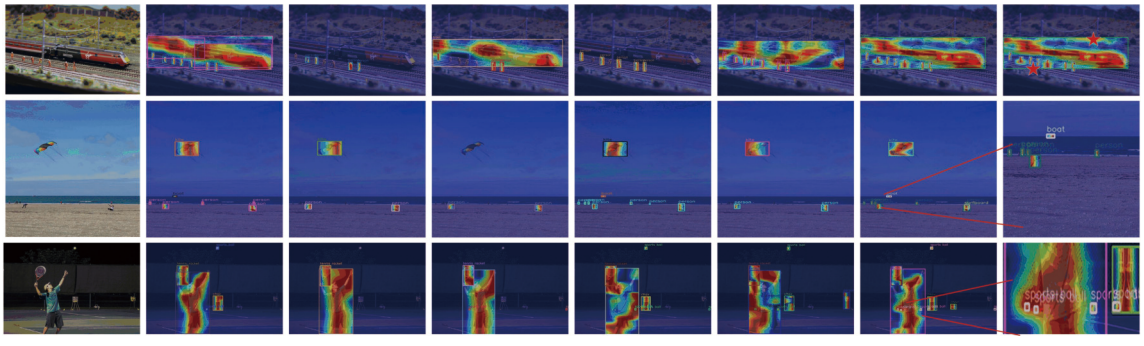
为了更直观验证本文提出的算法对于小目标检测的有效性，本文选取了 MS COCO 2017 验证集的图片分别进行预测检测框的可视化和热力图的可视化，如图 10 和图 11 所示。在图 10 中，在第 3 行的牛和飞鸟的检测中对比方法都有着不同程度的漏检及误检，而本文算法能更精确地检测目标，灵敏度更高、鲁棒性更强，表明本文算法对于小目标检测有着明显的提升。在图 11 中，对比算法在检测目标时热力图与检测目标形状有一定程度的偏差，而 Faster R-CNN 上还存在着重复检测的问题，如图 11 局部放大中所示，第 3 行在近处的人及远处的球同时在一张图像上且球又属于小目标时，各类算法都倾向把远处的球当作背景，而本文方法准确地识别了人和球都是检测目标。表明本文方法有效地区分前景和背景，改善了小标检测方面的问题，达到了更鲁棒的检测结果。

图 9　不同场景下测试图结果



a. FasterRCNN　b. FCOS　c. RetinaNet　d. Yolov3　e. YoloX　f. Ours

图 10　检测效果可视化对比



a. Input　b. FasterRCNN　c. FCOS　d. RetinaNet　e. Yolov3　f. YoloX　g. Ours　h. Partial Zoom

图 11　检测热力图可视化对比

4.5 消融实验

为了证明本文提出的颈部特征融合方法优于多尺度特征金字塔（FPN），本文将各个模型的颈部模块替换为本文提出的颈部模块后分别对比原始 FastRCNN-FPN、FCOS、RetinaNet。本文使用 COCO2017mini Val 作为本次消融实验的测试集，在表 6 中结果验证了本文提出的 Tri-Neck 模型具有更优的效果，分别超过了原始模型 1.6%、2.0%、2.4% 和 3.1%。

为了进一步验证本文提出的激活函数和优化交并比损失函数的有效性，首先使用 Tanh 和 IoU 的模型在 COCO2017mini Val 的实验作为基准，然后依次替换其中的函数，如表 6 所示。当仅增加原始的特征金子塔模块时，Faster R-CNN 的 mAP 和 mAPs 分别为 37.4% 和 21.2%，FCOS 的 mAP 和 mAPs 分别为 36.6% 和 21.0%，RetinaNet 的 mAP 和 mAPs 分别为 36.5% 和 20.4%，本文方法的 mAP 和 mAPs 分别为 37.0% 和 20.8%。其次，将特征金字塔调换为改进后的多尺度特征融合模块后分别替换激活函数和交并比函数，其中仅替换激活函数后，Faster R-CNN 模型提升了 1.9% 和 1.6%，FCOS 提升了 1.3% 和 1.0%，RetinaNet 提升了 1.7% 和 1.3%，本文方法提升了 1.3% 和 1.1%。仅替换交并比函数后，Faster R-CNN 提升了 1.2% 和 0.7%，FCOS 提升了 0.3% 和 0.2%，RetinaNet 提升了 1.1% 和 0.5%，本文方法提升了 0.8% 和 0.7%。最后，激活函数和交并比函数都被替换后，Faster R-CNN 提升了 2.8% 和 2.7%，FCOS 提升了 2.0% 和 1.4%，RetinaNet 提升了 2.4% 和 2.2%，本文方法提升了 2.0% 和 2.4%。表 6 中的实验结果表明，本文方法使用动态阀函数和改进后的交并比损失函数对均值平均精度具有提升的效果。

表 6 消融实验

Model	Backbone	epoch	FPN	Tri-Neck	Valve	iIoU	mAP/%	mAP_S/%
Faster R-CNN	ResNet50	12	√	—	—	—	37.4	21.2
			—	√	√	—	39.3	22.8
			—	√	—	√	38.6	21.9
			—	√	√	√	39.0	23.2
FCOS	ResNet50	12	√	—	—	—	36.6	21.0
			—	√	√	—	37.9	22.0
			—	√	—	√	36.9	21.2
			—	√	√	√	38.6	22.4
RetinaNet	ResNet50	12	√	—	—	—	36.5	20.4
			—	√	√	—	38.2	21.7
			—	√	—	√	37.6	20.9
			—	√	√	√	38.9	22.6
Ours	ResNet50	12	√	—	—	—	37.0	20.8
			—	√	√	—	38.3	21.9
			—	√	—	√	37.8	21.5
			—	√	√	√	**40.1**	**23.9**

5 结束语

本文提出了一种动态自适应通道注意力的小目标检测方法，首先提出一种 Tri-Neck 网络结构，采用同层级特征图的横向连接和相邻层级的下采样连接以进行特征融合，引入动态通道注意力模块来处理两种连接，从而增强前景和背景的对比度，进一步提高多尺度特征表示的能力。此外，本文还提出了新的动态通道注意力激活函数和改进后的交并比损失函数。这些改进方法能够有效提升模型性能，特别是在目标检测中受到背景噪声干扰时，模型能够保持稳定的检测结果。

本文的主要贡献如下。

1）提出一种 Tri-Neck 网络结构，以解决小目标过度抑制和特征缺失问题。

2）提出一种分组批量动态通道注意力模块，同时考虑通道之间的关系和位置信息，通过选择加权特征通道提升模型表达和泛化能力。

3）提出新的激活函数和交并比损失函数，增强有用信息通道、抑制冗余信息通道、提高了预测框的匹配性。

目前，基于 Transformer 作为骨干网络的检测模型已经达到了优秀效果，然而由于其多头注意力

机制，将带来大量的计算资源的消耗。为了解决该问题，未来将在 Transformer 多尺度融合及其轻量化方向作研究。

参 考 文 献

[1] 朱威, 王立凯, 靳作宝, 等. 引入注意力机制的轻量级小目标检测网络[J]. 光学 精密工程, 2022, 30(8): 998-1010.
ZHU W, WANG L K, JIN Z B, et al. Lightweight small object detection network with attention mechanism[J]. Optics and Precision Engineering, 2022, 30(8): 998-1010.

[2] 童小钟, 魏俊宇, 苏绍璟, 等. 融合注意力和多尺度特征的典型水面小目标检测[J]. 仪器仪表学报, 2023, 44(1): 212-222.
TONG X Z, WEI J Y, SU S J, et al. Typical small target detection on water surfaces fusing attention and multi-scale features[J]. Chinese Journal of Scientific Instrument, 2023, 44(1): 212-222.

[3] 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.

[4] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 6517-6525.

[5] REDMON J, FARHADI A. YOLOv3: An incremental improvement[EB/OL]. (2018-04-08)[2023-09-24]. https://arxiv.org/abs/1804.02767.

[6] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot MultiBox detector[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016: 21-37.

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

[8] DAI X Y, CHEN Y P, XIAO B, et al. Dynamic head: Unifying object detection heads with attentions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2021: 7369-7378.

[9] GAO S H, CHENG M M, ZHAO K, et al. Res2Net: A new multi-scale backbone architecture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(2): 652-662.

[10] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 580-587.

[11] LAW H, DENG J. Cornernet: Detecting objects as paired keypoints[C]//2018 Proceedings of the European Conference on Computer Vision. Munich: Springer, 2018: 734-750.

[12] 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. New York: IEEE, 2019: 6568-6577.

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

[14] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 2818-2826.

[15] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 1-9.

[16] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[EB/OL]. (2015-11-23)[2023-09-9]. https://arxiv.org/abs/1511.07122.

[17] HE K, ZHANG X, REN S, 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.

[18] 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. New York: IEEE, 2017: 936-944.

[19] WANG J, CHEN K, XU R, et al. Carafe: Content-aware reassembly of features[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 3007-3016.

[20] GUO C X, FAN B, ZHANG Q, et al. AugFPN: Improving multi-scale feature learning for object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 12592-12601.

[21] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770-778.

[22] HOWARD A G, ZHU M L, CHEN B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications[EB/OL]. (2017-04-17)[2023-01-25]. https://arxiv.org/abs/1704.04861.

[23] WU Y, HE K. Group normalization[J]. International Journal of Computer Vision, 2020, 128(3): 742-755.

[24] ZHANG R. Making convolutional networks shift-invariant again[C]//2019 International Conference on Machine Learning (ICML). Los Angeles: PMLR, 2019: 7324-7334.

[25] LI J, LYU S, LI Z. Unsupervised domain adaptation via

softmax-based prototype construction and adaptation[J]. Information Sciences, 2022, 609: 257-275.

[26] LI Y W, SONG L, CHEN Y K, et al. Learning dynamic routing for semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 8550-8559.

[27] YU J, JIANG Y, WANG Z, et al. Unitbox: An advanced object detection network[C]//Proceedings of the 24th ACM international conference on Multimedia. New York: ACM, 2016: 516-520.

[28] REZATOFIGHI H, TSOI N, GWAK J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]//2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE Press, 2019: 658-666.

[29] YAMADA Y, OTANI M. Does robustness on ImageNet transfer to downstream tasks?[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2022: 9205-9214.

[30] JAIN S, DASH S, DEORARI R. Object detection using Coco dataset[C]//2022 International Conference on Cyber Resilience (ICCR). Guangzhou: IEEE, 2022: 1-4.

[31] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE, 2017: 2999-3007.

[32] GE Z, LIU S T, WANG F, et al. YOLOX: Exceeding YOLO series in 2021[EB/OL]. (2021-07-18)[2023-09-25]. https://arxiv.org/abs/2107.08430.

[33] YANG Z, LIU S H, HU H, et al. RepPoints: Point set representation for object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2019: 9656-9665.

[34] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 9756-9765.

[35] ZHU X Z, SU W J, LU L W, et al. Deformable DETR: Deformable transformers for end-to-end object detection[EB/OL]. (2020-10-08)[2023-09-13]. https://arxiv.org/abs/2010.04159.

[36] SUN P Z, ZHANG R F, JIANG Y, et al. Sparse R-CNN: End-to-end object detection with learnable proposals[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2021: 14449-14458.

[37] XU C, WANG J W, YANG W, et al. RFLA: Gaussian receptive field based label assignment for tiny object detection[C]//2022 European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 526-543.

[38] CHENG G, YUAN X, YAO X W, et al. Towards large-scale small object detection: Survey and benchmarks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(11): 13467-13488.

编 辑 张 莉